NIST Special Publication
NIST SP 1283

NIST Conference Papers Fiscal Year 2021

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1283

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

NIST Special Publication
NIST SP 1283

NIST Conference Papers Fiscal Year 2021

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.1283

August 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 1283
August 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

NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2022-08-24

How to Cite this NIST Technical Series Publication
National Institute of Standards and Technology (2022) NIST Conference Papers Fiscal Year 2021. (U.S. Department of Commerce, Washington, D.C.), NIST Special Publication (SP) NIST SP 1283. https://doi.org/10.6028/NIST.SP.1283

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 2021 (October 1, 2020–September 30, 2021).

Keywords

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

Table of Contents

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 2021.

More information on public access to NIST research is available.

Experimental Comparison of Performance Monitoring Using Neural Networks Trained with Parameters Derived from Delay-Tap Plots and Eye Diagrams

Xiaoxia Wu[1], Jeffrey A. Jargon[2], Chih-Ming Wang[2], Loukas Paraschis[3], and Alan E. Willner[1]
1. Dept. of Electrical Engineering - Systems, University of Southern California, Email: xiaoxia@usc.edu
2. National Institute of Standards and Technology, Boulder, CO 80305 USA
3. Optical Networking, Advanced Technology and Planning, Cisco Systems, Inc.

Abstract: We experimentally demonstrate the use of artificial neural networks trained with parameters derived from both delay-tap plots and eye diagrams for multi-impairment monitoring in a 40-Gbit/s non-return-to-zero on-off keying system.

OCIS codes: (060.2330) Fiber optics communications; (100.4996) Pattern recognition, neural networks.

1. Introduction

As data rates increase and network architectures become more complex, it becomes more difficult to predict and manage data impairments due to degradations that can change with time. In order to enable robust and cost-effective "self-managed" operations, optical networks will need to be able to agilely monitor their physical states and the quality of propagating data signals, automatically diagnose and repair problems, redirect traffic, and dynamically allocate resources. Thus, optical performance monitoring (OPM) and automatic system control are becoming increasingly important [1, 2]. Key features of optical performance monitors are simplicity in implementation and the ability to accommodate different modulation formats and impairments.

OPM can be performed by measuring changes to the data and determining "real-time" changes resulting from various impairments, such that a change in a particular effect will alter a measured parameter. This can employ: (i) optical techniques to monitor changes in a radio frequency (RF) tone power or in the spectral channel power distribution [3], or (ii) electrical post-processing techniques in the specific case of coherent detection [4]. The optical approaches have been shown to be powerful for OPM. However, the electrical distortions that are crucial for the signal quality at the decision point tend to be neglected in the optical approaches. Techniques proposed for OPM using off-line digital signal processing of received electrical data signals include the use of (i) amplitude histograms, power distributions, and asynchronous sampling to estimate bit error rate (BER) [5-8]; (ii) delay-tap plots to distinguish among impairments [9-12]; and (iii) pattern recognition techniques to identify multiple impairments [13, 14]. Recently, we have proposed a neural network approach to train receivers in an optical network to distinguish among resultant shapes of either the data channel's eye diagrams or asynchronous delay-tap plots in the presence of the degrading effects of optical signal-to-noise ratio (OSNR), chromatic dispersion (CD), and polarization-mode dispersion (PMD) [15, 16].

In this paper, we experimentally compare the use of artificial neural networks (ANNs) trained with parameters derived from both delay-tap plots and eye diagrams to simultaneously identify OSNR, CD and PMD in a 40-Gbit/s non-return-to-zero on-off keying (NRZ-OOK) system. The monitoring range is 18-30 dB for OSNR, 0-100 ps/nm for CD, and 0-10 ps for differential group delay (DGD), i.e. first order PMD. A correlation coefficient of 0.995 is obtained when using delay-tap plots. This method exhibits slightly better performance compared with using eye diagrams, where a correlation coefficient of 0.972 is achieved.

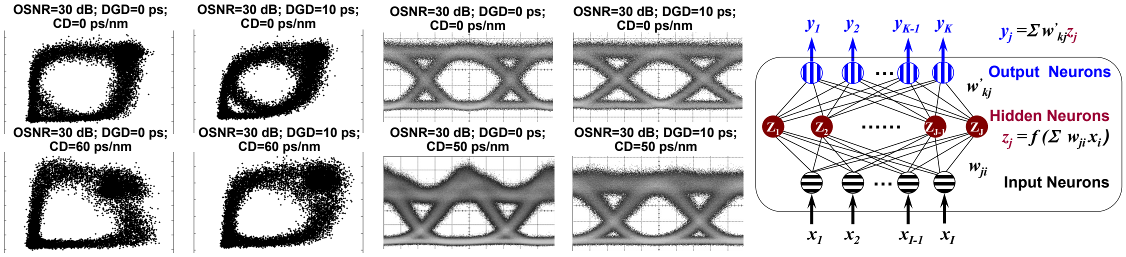
2. Concept

With asynchronous delay-tap sampling, each sample point is comprised of two measurements separated by a specific time corresponding to the length of the delay [9]. Fig. 1 (a) illustrates one-half bit-period (B/2) delay-tap plots for a 40-Gbit/s NRZ-OOK signal at a few select combinations of OSNR, CD and DGD. Fig. 1 (b) shows the eye diagrams with slightly different combinations of impairments. Visually, it is obvious that these impairments produce distinct features in both the delay-tap plots and eye diagrams.

To simultaneously quantify the impairments, we use ANNs trained with parameters derived from delay-tap plots or eye diagrams. ANNs are information-processing systems that learn from observations and generalize by abstraction, which consist of multiple layers of processing elements called neurons [17]. Each neuron is linked to other neurons in neighboring layers by varying coefficients, as shown in Fig. 1 (c). ANNs learn the relationships among sets of input-output data that are characteristic of the device or system under consideration. After the input

Jargon, Jeffrey; Wu, Xiaoxia; Wang, Chih-Ming; Willner, Alan. "Experimental comparison of performance monitoring using neural networks trained with parameters derived from delay-tap plots and eye diagrams." Presented at Optical Fiber Communication Conference (OFC), San Diego, CA, US. March 21, 2010 - March 25, 2010.

vectors are presented to the input neurons and output vectors are computed, the ANN outputs are compared to the desired outputs and errors are calculated. Error derivatives are then calculated and summed for each weight until all of the training sets have been presented to the network. The error derivatives are used to update the weights for the neurons, and training continues until the errors reach prescribed values. After training, the ANN can be tested by use of other sets of data.



(a) Delay-tap plots with impairments (b) Eye diagrams with impairments (c) Artificial neural networks

Fig. 1: Concepts of delay-tap plots, eye diagrams, and artificial neural networks.

3. Experiment Setup and ANN Models

Fig. 2 shows the experimental setup. The 40-Gbit/s NRZ-OOK signal is generated using a Mach-Zehnder modulator (MZM), driven by a 40-Gbit/s pseudo-random bit sequence (PRBS). The signal then goes through a DGD emulator, followed by a tunable dispersion compensating module (TDCM), which serves as the CD emulator. The DGD emulator has a range of -50 to +250 ps with a resolution of 0.002 ps, and the TDCM has a tuning range of +/- 400 ps/nm and a 10 ps/nm tuning resolution. The output of the TDCM is sent to an erbium-doped fiber amplifier (EDFA) with a variable optical attenuator (VOA) in front to adjust the received OSNR. The noise-loaded signal is then filtered by a bandpass filter (BPF) with 1 nm bandwidth, and sent to a sampling oscilloscope, where the waveform of the signal is recorded. The delay-tap asynchronous plots and eye diagrams can then be constructed prior to the parameters being extracted. In our experiment, we vary OSNR, CD and DGD to get a set of 252 waveforms (OSNR (dB): 18-30 in steps of 2; CD (ps/nm): 0-100 in steps of 20; DGD (ps): 0-10 in steps of 2), of which 137 combinations are used for training and 115 combinations are used for testing.

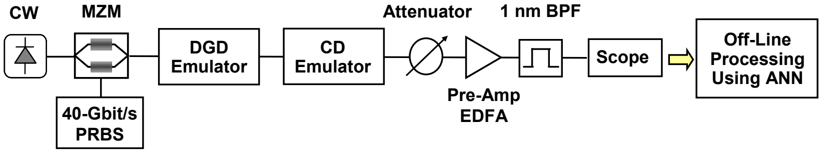


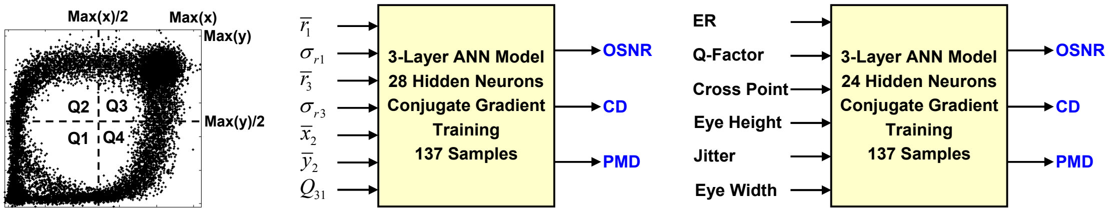
Fig. 2: Experimental setup. CW: continuous wave; MZM: Mach-Zehnder modulator; BPF: bandpass filter.

To quantify the distinct features, we derive parameters from the diagrams. For eye diagrams, we utilize widely-used parameters such as extinction ratio, Q-factor, crossing amplitude, eye height, jitter, and eye width [18]. For delay-tap plots, we divide the plots into four quadrants Q1-Q4 [16]. The data pairs are divided into the quadrants as follows: $(x_i, y_i) \in \text{Q1}$ if $\{0 \le x_i \le \text{Max}(x)/2$ and $0 \le y_i \le \text{Max}(y)/2\}$; $(x_i, y_i) \in \text{Q2}$ if $\{0 \le x_i \le \text{Max}(x)/2$ and $\text{Max}(y)/2 < y_i \le \text{Max}(y)\}$; $(x_i, y_i) \in \text{Q3}$ if $\{\text{Max}(x)/2 < x_i \le \text{Max}(x)$ and $\text{Max}(y)/2 < y_i \le \text{Max}(y)\}$; and quadrant 4 is not used since it contains data that is the mirror image of quadrant 2. Fig. 3 (a) illustrates this concept. With three quadrants defined, we can perform some basic statistical calculations on the data within each quadrant, such as means and standard deviations. For quadrants 1 and 3, we calculate the means and standard deviations of the magnitudes $(\bar{r}_1, \sigma_{r1}, \bar{r}_3, \sigma_{r3})$. For quadrant 2, we calculate the means and standard deviations of the x's and y's separately since this quadrant is on the off-diagonal. For the purpose of training our ANNs, we do not make use of the second quadrant's standard deviations since they do not vary significantly with different combinations of impairments. One final parameter we make use of is similar to the Q-factor, which we define as $Q_{31} = (\bar{r}_3 - \bar{r}_1)/(\sigma_{r1} + \sigma_{r3})$. The ANN using delay-tap parameters consists of seven inputs $(\bar{r}_1, \sigma_{r1}, \bar{r}_3, \sigma_{r3}, \bar{x}_2, \bar{y}_2, Q_{31})$, three outputs (OSNR, CD, and DGD), and 28 hidden neurons, which is illustrated in Fig. 3 (b). Similarly, a block diagram for the ANN using eye diagram parameters is shown in Fig. 3 (c).

The training errors of the ANNs were 0.0309 and 0.1147 for the case of using parameters derived form delay-tap plots and eye diagrams, respectively. The ANNs were trained by use of a software package developed by Zhang et al. [19]. Although alternatives were explored, a conjugate-gradient technique was chosen since it offers a nice

Jargon, Jeffrey; Wu, Xiaoxia; Wang, Chih-Ming; Willner, Alan. "Experimental comparison of performance monitoring using neural networks trained with parameters derived from delay-tap plots and eye diagrams." Presented at Optical Fiber Communication Conference (OFC), San Diego, CA, US. March 21, 2010 - March 25, 2010.

compromise in terms of memory requirements and implementation effort.

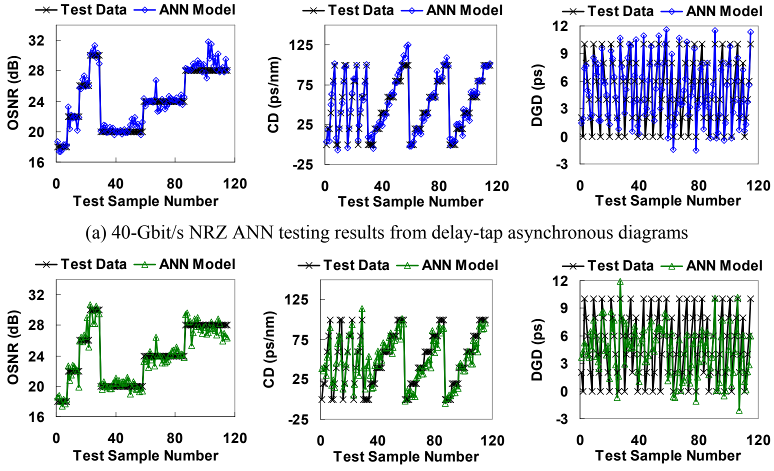


(a) Parameter derivation (b) ANN with parameters from delay-tap plots (b) ANN with parameters from eye diagrams

Fig. 3: Parameter derivation and ANN block diagrams. ER: extinction ration.

4. Experimental Results and Discussions

Once the models were trained, we validated their accuracies with a different set of 115 combinations as testing data. The software reported a correlation coefficient of 0.995 when using the parameters derived from delay-tap plots and 0.972 in the case of using eye-diagram parameters. The results are shown in Figs. 4 (a) and (b), respectively. When parameters from delay-tap plots were used, the root-mean-square (RMS) errors were 0.919 dB for OSNR, 6.368 ps/nm for CD, and 1.479 ps for DGD; when the parameters from eye diagrams were used, the RMS errors were 0.866 dB for OSNR, 14.642 ps/nm for CD, and 2.479 ps for DGD. In this particular case, results were slightly better when using parameters from the delay-tap plots. Fig. 4 (c) summarizes and compares the two cases.



(a) 40-Gbit/s NRZ ANN testing results from delay-tap asynchronous diagrams

(b) 40-Gbit/s NRZ ANN testing results from eye diagrams

	Delay-Tap Asynchronous Diagrams	Eye Diagrams
Training Error	0.0309	0.1147
Training Samples	137	115
Testing Samples	137	115
Correlation Coefficient	0.995	0.972
OSNR RMS Error	0.919 dB	0.866 dB
CD RMS Error	6.368 ps/nm	14.642 ps/nm
DGD RMS Error	1.479 ps	2.479 ps

(c) Comparison of using delay-tap plots and eye diagrams.

Fig. 4: Experimental results.

5. Conclusions

We have shown that ANN models trained with parameters derived from both measured delay-tap plots and eye diagrams can effectively be used to simultaneously identify levels of OSNR, CD, and DGD for 40-Gbit/s NRZ-OOK signals.

This work is supported by the U.S. Department of Commerce and the DARPA CORONET program, and is not subject to U.S. Copyright. The authors thank Jeng-Yuan Yang and Xue Wang from USC for their help.

6. References

[1] D. C. Kilper et al, *JLT* 22 (1), 294-304 (2004).
[2] Y. C. Chung, *ECOC '08*, paper We.1.D.1.
[3] T. Luo et al, *OFC'03*, paper ThY3.
[4] H. Sun et al, *Opt. Express* 16 (2), 873-879 (2008).
[5] I. Shake et al, *Electron. Lett.* 34 (22), 2152-2154 (1998).
[6] N. Hanik et al, *Electron. Lett.* 35 (5), 403-404 (1999).
[7] S. Ohteru et al, *PTL* 11 (10), 1307-1309 (1999).
[8] I. Shake et al, *JLT* 22 (5), 1296-1302 (2004).
[9] S. D. Dods et al, *OFC'06*, paper OThP5.
[10] S. D. Dods et al, *OFC'07*, paper OMM5.
[11] B. Kozicki et al, *Opt. Express* 16 (6), 3566-3576 (2008).
[12] B. Kozicki et al, J. Opt. Netw. 6 (11), 1257-1269 (2007).
[13] R. A. Skoog et al, *PTL* 18 (22), 2398-2400 (2006).
[14] T. B. Anderson et al, *JLT* 27 (16), 3729-3736 (2009).
[15] X. Wu et al, *JLT* 27 (16), 3580-3589 (2009).
[16] J. A. Jargon et al, *OFC'09*, paper OThH1.
[17] M. H. Hassoun, the MIT Press (1995).
[18] J. A. Jargon et al, *JLT* 26 (21) 3592-3600 (2008).
[19] "NeuroModeler, ver. 1.5," Q. J. Zhang, et al (2004).

Jargon, Jeffrey; Wu, Xiaoxia; Wang, Chih-Ming; Willner, Alan. "Experimental comparison of performance monitoring using neural networks trained with parameters derived from delay-tap plots and eye diagrams." Presented at Optical Fiber Communication Conference (OFC), San Diego, CA, US. March 21, 2010 - March 25, 2010.

Optimal Time of Use of Renewable Electricity Pricing: Three-Player Games Model

Siham KHOUSSI*†, Hasnae BILIL*†, Ghassane ANIBA*
* Mohammadia School of Engineers, Mohammed V University of Rabat, Morocco
ghassane@emi.ac.ma
† Advanced Network Technologies Division, National Institute of Standards and Technology
{siham.khoussi,hasnae.bilil}@nist.gov

***Abstract*—Currently, the electricity demand is exponentially increasing due to the population growth. Therefore, the demand side management (DSM) is becoming unavoidable especially with the increasing use of renewable energy sources. One of the most known tactics of DSM is the use pricing strategies to threaten users to schedule their loads by controlling their own appliances. In this paper, a new model of electricity market operators is proposed based on three actors: the utility grid (G) with renewable energy (RE) generation, the electricity consumer (U) and a storage company (S). This approach aims to develop the adequate hourly prices which optimize the utility function of each operator. Then, in order to deal with this objective, two related games are defined. The first one is based on the satisfaction function of U and the G and aims to give the hourly prices of consumers' electricity bills. While the second one is based on the satisfaction function of G and S in order to optimize the hourly prices of G's electricity bills. Finally, a case study based on a given RE production and consumers load forecasts has been considered. Simulation results show that the obtained hourly prices allows the consumer load curve to follow the RE curve generation while achieving the main objective of the proposed approach.**

***Index Terms*—Renewable Energy based generation, optimization, game theory three players, outsourced storage.**

I. Introduction

For a long tome, the utility grid has been providing users with electricity for their appliances to function well. Nevertheless, it's still having problems with the load curves' daily fluctuations. In fact, the load curve usually presents peak hours, semi-peaks, and off-peak hours. This unregulated variations are very uncomfortable for the utility companies, since they have to turn on and off some of their massive generators to meet the necessary demand which might result in a huge loss in generation capacities in case of traditional production. Additionally, in the case of renewable energy based production, companies usually seek a perfect consumption, that is when users use up all what has been generated by the suppliers. Therfore, they often seek to keep the curve within a specific pattern in order to avoid any technical problems that might harm their equipements. Demand side management is when the utility grid interferes with users' demand to regulate the bumps on the load curve and get a steady curve eventually [1]. It is done through many measures [2]. A famous technique is the use of pricing to motivate users to lower or increase their loads in times of peak or off-peak hours. This is called time of use pricing [3], [4], it's when suppliers set their electricity prices in advance based on users' response and/or customer's satisfaction.

In this paper, we focus on renewable sources of energy (RE) as the main source of generation. We also assume that we only have a single generation company or that they are all grouped together under one company G. Furthermore, we suppose that all storage services have been completely externalized to storage companies. Likewise, we gather them all under one company S. The same goes for users, we use U to indicate all users of all kinds. We assume that each of the above participants stick to their assigned tasks. That is, users' unique task is to consume energy unlike [5] where they're participating in both storage and generation. The generation company G satisfies users' demand by all means. And, the storage company's single task is to purchase electricity from G when they have excess and sell it back when G needs it again. All these assumptions have been done primarily to simplify our modeling process. Nevertheless, this model can still be extended into different scenarios, like for instance when each player participates in other tasks not just the one we assigned to him. One more thing to point out is that we use game theory to optimize the utility functions of G, S and U because of their usual selfishness and eagerness to win. Thereby, we ought to call the interactions between them games and the actors players to emphasize the fact that each of them will only play winning strategies.

The rest of this paper is organized like this: Section II models all players roles and defines their tasks as well as their utility functions, Section III contains our derivation of the equilibrium, Section IV supports our model with a study case including its performance analysis and Section V concludes this paper.

II. Presenting the proposed model

A. Defining the players' roles and their utility functions

In this section, the three players G, S and U play in two different games. Game 1, between G and U and Game 2 between G and S, are shown in Fig.1. First, in Game 1, users

This work is realised in the framework of a research project (reference: InnoTherm-13-MicroCSP), funded by The Institute of Research in Solar Energy and New Energies (IRESEN), Morocco.

Khoussi, Siham. "Optimal Time of Use of Renewable Electricity Pricing: Three-Player Games Model." Presented at International Conference on Smart Grid Communications (IEEE SmartGridComm 2015), Miami, FL, US. November 02, 2015 - November 05, 2015.

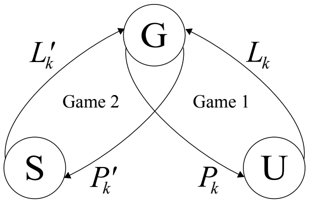


Fig. 1: Three-Player Games Model

buy electricity from G with the price P_k per unit to satisfy their load L_k, but given that G has no control over its active power generation g_k as it uses renewable sources of energy which are unreliable because they often rely on the weather, we obtain situations where G also needs to purchase or sell the difference between the generation and users' load L_k to storage owners S with the price P'_k per unit, this is Game 2. Therefore, on one hand we model G as a supplier for users in Game 1 and as a client of S in Game 2. Next, we associate each player U, G and S with a utility function. Let U_u, U_g and U_s be their utility functions respectively.

To begin with, we define two functions s and s'. The first one denotes users' satisfaction towards the price P_k by comparing the load L_k with the nominal users' demand d_k. If $d_k > L_k$ then users aren't satisfied with the price P_k. But when $d_k < L_k$, it's completely the opposite since users find P_k suitable to the extent of increasing their load above the nominal value d_k.

In Game 2, G duplicates users' actions in Game 1 as it acts as a client to S, the second function s'_k, has exactly the same signification of s_k but this time we consider that G has a load L'_k needed to meet users' demand of energy in the cases of $g_k < L_k$, $g_k = L_k$ and $g_k > L_k$. This means that L'_k could be either positive, null or negative depending on the situation of G either having a surplus or is in need. And the same goes for P'_k, the price by which S sells/purchases L'_k to/from G. Although, it seems that G is always the buyer and S is always the seller in Game 2, it all depends on the users in Game 1. In fact, if users ask for more than what G has, S provides that difference buy selling it to G. And if users' load is below the generation curve, G purchases a negative load from S. This is what we mean by L'_k not always being positive.
Let $d'_k = d_k - g_k$ be the nominal value that G needs to satisfy all users' nominal demand.

In the rest of this paper, we will be using the same satisfaction function that has been used in [6] to represent our model. The functions s_k and s'_k both must first satisfy the following conditions:

- For s_k:
  1) If $d_k = L_k$, $\quad s_k(L_k, d_k) = 0$
  2) $L_k > d_k$
  $$\frac{\partial s_k}{\partial L_k} < 0,\ \frac{\partial^2 s_k}{\partial L_k^2} > 0$$
  3) $L_k < d_k$
  $$\frac{\partial s_k}{\partial L_k} < 0,\ \frac{\partial^2 s_k}{\partial L_k^2} > 0$$
- For s'_k:
  1) If $d'_k = L'_k$, $\quad s'_k(L'_k, d'_k) = 0$
  2) $L'_k > d'_k$
  $$\frac{\partial s'_k}{\partial L'_k} < 0,\ \frac{\partial^2 s'_k}{\partial L'^2_k} > 0$$
  3) $L'_k < d'_k$
  $$\frac{\partial s'_k}{\partial L'_k} < 0,\ \frac{\partial^2 s'_k}{\partial L'^2_k} > 0$$

The following two functions respect the above conditions [6].

$$\begin{aligned} s_k(L_k, d_k) &= d_k\beta_k\left(\left(\frac{L_k}{d_k}\right)^{\alpha_k} - 1\right) \\ s_k(L'_k, d'_k) &= d'_k\beta'_k\left(\left(\frac{L'_k}{d'_k}\right)^{\alpha'_k} - 1\right) \end{aligned} \tag{1}$$

It is important to realize that as long as $\alpha_k < 0$, $\alpha'_k < 0$ and $\alpha_k\beta_k < 0$, $\alpha'_k\beta'_k < 0$, our model can be modified while still having the same results. Fig. 2 shows a simple example of how s_k and s'_k could look like.

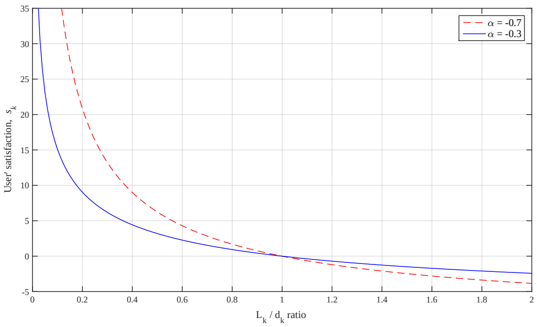


Fig. 2: Satisfaction function with two alpha values $\alpha = -0.7$, $\alpha = -0.3$ and $\beta = 5$

Finally, we define all players' objective functions to be their (electricity sales - purchases - costs - satisfaction function) since they are all concerned with minimizing their costs and maximizing their profits [7]. The same applies to users, we take out the electricity sales since in this scenario they are not participating in the production nor in the storage.

B. Problem formulation

This section is about developing utility functions of the players in each game. We divide a day into multiple time slots, $k \in \mathcal{N} = \{1, \ldots, N\}$.

First in Game 1, users buy electricity L_k from G with P_k per unit. Therefore, their purchases are P_kL_k. By subtracting their satisfaction function, we get the following utility function:

$$U_u = -\sum_{k=1}^{N}\left[P_kL_k + d_k\beta_k\left(\left(\frac{L_k}{d_k}\right)^{\alpha_k} - 1\right)\right] \tag{2}$$

Bear in mind that users' load can not exceed a certain value $L_{k,\max}$ which is normally the sum of all users' appliances turned on all at once, nor fall behind $L_{k,\min}$ which is the minimal load that users require from G. Under this condition, users' optimization problem becomes:

$$\begin{aligned} \max_{L_k} \quad & U_u \\ \text{subject to} \quad & L_{k,\min} \leq L_k \leq L_{k,max} \\ & L_{k,\max} = \min(d_{k,max}, (d'_{k,max} + g_{k,max})) \\ & L_k,\ \forall k \in \{1, 2, \dots, N\} \end{aligned} \tag{3}$$

Next, we said that G takes on two roles, a provider of electricity to users in Game 1 and a mere user to S in Game 2. Hence, his utility function consists of two parts U_1 and U_2 of the first and second roles correspondingly.

When G is supplying electricity in the first game, G earns $P_k L_k$. And since his source of energy is mainly renewable we consider his operating and maintenaning costs C_g instead of his production costs. Consequently, G's profit is $(P_k L_k - c_g g_k)$. We also take into consideration the costs proceeding the variations between the generation g_k and the average $\bar{g}$ multiplied by a factor μ [6], such that:

$$f(g) = \sum_{k=1}^{N} \mu (g_k - \bar{g})^2 \tag{4}$$

We then write:

$$U_1 = \sum_{k=1}^{N} \left[P_k L_k - c_g g_k - d_k \beta_k \left(\left(\frac{L_k}{d_k} \right)^{\alpha_k} - 1 \right) \right] \\ -f(g) \tag{5}$$

In the second situation, G purchases $P'_k L'_k$ from S, and expresses his satisfaction through s'. Thus, the latter half of his utility function U_2 is similar to users' objective function U_u:

$$U_1 = -\sum_{k=1}^{N} \left[P'_k L'_k + d'_k {\beta'}_k * \left(\left(\frac{L'_k}{d'_k} \right)^{\alpha'_k} - 1 \right) \right] \tag{6}$$

To obtain G's utility function we sum both U_1 and U_2 together:

$$\begin{aligned} U_g = & \sum_{k=1}^{N} \left[P_k L_k - c_g g_k - d_k \beta_k \left(\left(\frac{L_k}{d_k} \right)^{\alpha_k} - 1 \right) \right] \\ & -f(g) \ - \sum_{k=1}^{N} \left[P'_k L'_k + d'_k {\beta'}_k \left(\left(\frac{L'_k}{d'_k} \right)^{\alpha'_k} - 1 \right) \right] \end{aligned} \tag{7}$$

Therefore, the optimization problem of G becomes:

$$\begin{aligned} \max_{L'_k, P_k} \quad & U_g \\ \text{subject to} \quad & L'_{k,min} \leq L'_k \leq L'_{k,max}, \qquad c_g \leq P_k \\ & L'_{k,\max} = \min \left(d'_{k,max}, \frac{g'_{k,max}}{\phi} \right) \\ & L'_k, P_k \ ; \forall k \in \{1, 2, \dots, N\} \end{aligned} \tag{8}$$

The variable g' stands for storage owners' active generated power required by G. We add a constraint on g' to imply that S only sells/purchases what he receives from G on a timely basis no more no less, multiplied by a coeffecient:

$$g'_k = \left(\frac{1}{\eta_d} + \eta_c \right) L'_k \tag{9}$$

To simplify, we put :

$$\phi = c_s \left(\frac{1}{\eta_d} + \eta_c \right) \tag{10}$$

Whereas η_c and η_d are respectively the charging and discharging efficiencies of storage equipements owned by S and c_s is their operations and maintenance costs [8].

As for storage owners S, and by following the footsteps of G when he's acting as a supplier in Game 1, we get this utility function:

$$U_s = \sum_{k=1}^{N} \left[P'_k L'_k - c_s g'_k - d'_k * {\beta'}_k \left(\left(\frac{L'_k}{d'_k} \right)^{\alpha'_k} - 1 \right) \right] \\ -f(g') \tag{11}$$

So using (10) U_s becomes:

$$U_s = \sum_{k=1}^{N} \left[P'_k L'_k - \phi L'_k - d'_k \beta'_k \left(\left(\frac{L'_k}{d'_k} \right)^{\alpha'_k} - 1 \right) \right] - \\ f(g') \tag{12}$$

Again, $f(g')$ stands for the costs resulting of the variations between the delivered amounts and the average $\bar{g'}$ taking into account a coefficient μ'.

$$f(g') = \sum_{k=1}^{N} \mu' (g_k - \bar{g'})^2 \tag{13}$$

Thus, the storage company's optimization problem is:

$$\begin{aligned} \max_{P'_k} \quad & U_s \\ \text{subject to} \quad & P'_k \geq \phi \ \forall k \in \{1, 2, \dots, N\} \end{aligned} \tag{14}$$

Now that all players' utility functions have been set, we need to explain the rules of the games. In Game 1, G decides on the time of use price beforehand and users react upon that price. Identically in Game 2, S sets his pricing strategy and G responds by demanding a suitable load. Let $\mathcal{U}$, $\mathcal{G}$ and $\mathcal{S}$ be the strategy set of U, G and S respectively.

$$\mathcal{S} = \{P' | P' \in \mathrm{I\!R}^{\mathrm{N}}, L'\text{min} \leq L'(P') \leq L'\text{max}, \\ P' \geq \phi\} \tag{15}$$

$$\mathcal{G} = \{(P, L') | P, L' \in \mathrm{I\!R}^{\mathrm{N}}, L\text{min} \leq L(P) \leq L\text{max}, \\ P \geq cg, L'\text{min} \leq L' \leq L'\text{max}\} \tag{16}$$

$$\mathcal{U} = \{L | L \in \mathrm{I\!R}^{\mathrm{N}}, L\text{min} \leq L \leq L\text{max}\} \tag{17}$$

L and L' are set to be functions of P and P' because it's usually the prices that are being set first and determine how the loads will go.

Khoussi, Siham. "Optimal Time of Use of Renewable Electricity Pricing: Three-Player Games Model." Presented at International Conference on Smart Grid Communications (IEEE SmartGridComm 2015), Miami, FL, US. November 02, 2015 - November 05, 2015.

The purpose of the two games is to find a nash equilibrium for each [9], such that none of the players is motivated to change his stragtegy because no other choice has better payoffs than that one. We obtain a NASH equilibrium in Game 2 when:

$$\forall P' \in \mathcal{S}, P' \neq P'^{*} : U_s(P'^{*}, L'^{*}) \geq U_s(P', L'^{*}) \tag{18}$$

$$\forall L' \in \mathcal{G}_2, L' \neq L'^{*} : U_2(P'^{*}, L'^{*}) \geq U_2(P'^{*}, L') \tag{19}$$

Likewise, we get a nash equilibrium in Game 1 when:

$$\forall P \in \mathcal{G}_1, P \neq P^{*} : U_1(P^{*}, L^{*}) \geq U_1(P^{*}, L) \tag{20}$$

$$\forall L \in \mathcal{U}, L \neq L^{*} : U_u(P^{*}, L^{*}) \geq U_u(P^{*}, L) \tag{21}$$

III. Optimizing all three utility functions

In our proposed model, S takes action first by setting the electricity price P'_k per unit and G responds by adjusting his load L'_k. Next, G sets the price P_k followed by users who also accommodate their load L_k to P_k. Thus, we ought to start with Game 2 and then Game 1. Both games are multistage games. So, we will use backward induction to solve them [9].

A. Optimal demand reponse of G to the pricing strategy of S

By taking $P'_k, k \in \{1, 2, \ldots, N\}$ as given, we compute the first order derivative of U_g with respect to $L'_k, k \in \{1, 2, \ldots, N\}$.

$$\frac{\partial U_g}{\partial L'_k} = -P'_k - \alpha'_k \beta'_k \left(\frac{\partial L'_k}{\partial d'_k} \right)^{\alpha_k - 1} \tag{22}$$

When setting $\frac{\partial U_g}{\partial L'_k} = 0$, we obtain :

$$L'^{*}_k = d'_k \left(\frac{-P'_k}{\alpha'_k \beta'_k} \right)^{\frac{1}{\alpha'_k - 1}}, k \in \{1, 2, \ldots, N\} \tag{23}$$

Next, we need to compute the hessian matrix of U_g. The second order derivative of U_g is :

$$\frac{\partial^2 U_g}{\partial L'_k \partial L'_i} = \begin{cases} 0 & \text{if } k \neq i \\ \alpha'_k \beta'_k (\alpha'_k - 1) \dfrac{{L'_k}^{\alpha'_k - 2}}{{d'_k}^{\alpha'_k - 1}} & \text{if } k = i \end{cases} \tag{24}$$

The off-diagonal elements of the hessian matrix of U_g are all equal to zero and its diagonal elements are all negative since $\alpha_k < 1$ and $\alpha_k \beta_k < 0$. This implies that the solution L'_k that we found is indeed the local maximum.

B. Optimal pricing strategy of S based on the response of G

In the previous sub-section, we obtained the utility grid's optimal demand reponse to the storage company's price. In this section, we will be looking for the storage company's best pricing strategy that maximizes its utility function by replacing L'_k with L'^{*}_k in the utility function of S.

$$U_s = \sum_{k=1}^{N} \left[P'_k L'^{*}_k - \phi L'^{*}_k - d'_k {\beta'}_k \left(\left(\frac{L'^{*}_k}{d'_k} \right)^{\alpha'_k} - 1 \right) \right] - f(g') \tag{25}$$

The constraints on the sides of G define also the constraints on the price P'_k, that is:

$$P'_{k,\min} \leq P'_k \leq P'_{k,max} \tag{26}$$

Whereas

$$P'_{k,min} = \max \left(\phi, -\alpha'_k \beta'_k \left(\frac{L'_{k,max}}{d'_k} \right)^{\alpha_k - 1} \right)$$

and

$$P'_{k,\max} = -\alpha'_k \beta'_k \left(\frac{L'_{k,\min}}{d'_k} \right)^{\alpha_k - 1}$$

Thereby, the second optimization problem is:

$$\begin{array}{ll} \max\limits_{P} & U_s \\ \text{subject to} & P'_{k,min} \leq P'_k \leq P'_{k,max} \quad \phi \leq P'_k \\ & P'_k, \ \forall k \in \{1, 2, \ldots, N\} \end{array} \tag{27}$$

This problem is non linear with linear constraints, it could be resolved either by using an optimization software. To insure the existence of such optimum, we need to prove the negative definiteness of the hessian matrix which in this situation is parameter dependent. This has been done in [6].

C. Users' optimal demand reponse to the utility company's prices

This section is similar to section III.A. We replace the previously mentioned parameters of Game 2 by those of Game 1. We know that G plays first by setting his selling price per unit P_k and then users regulate their load L_k accordingly. Therefore, since also Game 1 is a multistage game, once again we use backward induction that dictates that we move backwards from L_k to P_k.

Hence, Users optimal demand response to the utility company's prices is found by following the same steps as before. Given that the utility function of users is:

$$U_u = -\sum_{k=1}^{N} \left[P_k L_k + d_k \beta_k \left(\left(\frac{L_k}{d_k} \right)^{\alpha_k} - 1 \right) \right] \tag{28}$$

When setting $\frac{\partial U_u}{\partial L_k} = 0$, we get :

$$L^{*}_k = d_k (\frac{-P_k}{\alpha_k \beta_k})^{\frac{1}{\alpha_k - 1}} \tag{29}$$

Again, the second order derivative is as follows:

$$\frac{\partial^2 U_u}{\partial L_k \partial L_i} = \begin{cases} 0 & \text{if } k \neq i \\ \alpha_k \beta_k (\alpha_k - 1) \dfrac{{L_k}^{\alpha_k - 2}}{{d_k}^{\alpha_k - 1}} & \text{if } k = i \end{cases} \tag{30}$$

Based on our choice of α_k and β_k ($\alpha_k < 1$ and $\alpha_k \beta_k < 0$), we say that $L_k, k \in \{1, 2, \ldots, N\}$is the optimal load for users givenP_k.

D. Optimal pricing based on users response

Now, we should find the optimum pricing strategy P^{*}_k that G must adopt to maximize his utility function with respect to L^{*}_k. But instead of going through the same demonstrations again, we use power balance between users' demand and G's demand. So, P^{*}_k is limited by on one hand L^{*}_k found in Game 2 and on the other hand by L^{*}_k of Game 1. Though G hopes that the difference between users demand and his own generation does not surpass nor fall behind his own optimal load L'^*_k that:

$$L'^*_k(P'^*_k) = L^*_k(P^*_k) - g_k \tag{31}$$

We can write the following equation:

$$d'_k \left(\frac{-P'^*_k}{\alpha_k \beta_k}\right)^{\frac{1}{\alpha_k - 1}} = d_k \left(\frac{-P^*_k}{\alpha_k \beta_k}\right)^{\frac{1}{\alpha_k - 1}} - g_k \tag{32}$$

By solving this equation we get:

$$P_k = -\alpha_k \beta_k \left(\frac{d'_k \left(\frac{-P'^*_k}{\alpha'_k \beta_k}\right)^{\frac{1}{\alpha_k - 1}} - g_k}{d_k} \right)^{\alpha_k - 1} \tag{33}$$

The interpretation of this formula is that when G is setting his selling priceP^*_k, he also takes in consideration Ss selling price P'^*_k, the nominal load values d_k and d'_k and the generation g_k, $k \in \{1, 2, \ldots, N\}$.

IV. Performance analysis

In this example we will be using wind as the only available source of energy. Nevertheless, this model can still be adjusted to contain different sources or a mix of different renewable sources of energy (RE). We collected data of the daily active power generation and users nominal and actual loads from [10] and [11]. As for the fixed costs of operating and maintaining a renewable power generation plant, it is generally said that they decrease as the size of the project increases [12] and it is not much compared to the costs of producing energy with traditional sources. The parameter $e = \frac{1}{(\alpha_k - 1)}$ is similar to the price elasticity [13].

Since in this example we are collecting data of Ameren Illinois of over 11 000 residential we will be using their elasticity values. n and n' could be considered as the nominal prices corresponding to the nominal demand d and d'. This has been done to simplify our demonstration. We take n and n' to be both 2.5. However, each of these parameters should be considered carefully before choosing appropriate values [14], [15]. Last but not least, we go with $u = 1$ and $u = 1$. For storage owners, [16]–[18] present a cost analysis of different energy storage technologies.

To begin with, in game two, G acts as a client towards S. Fig .3a shows the prices that S must set in order to optimize its utility function while taking in consideration the buyer's reaction G. We notice from Fig.4b that when the price is set too high or too low, the actual load curve happens to fall behind or surpass the nominal demand. This leads to an increase or decrease in the satisfaction function of G (s') and adds up to its objective function. By comparing the actual load curve to the nominal curve, we can clearly see that it's almost constant and flat. In fact, there are no longer huge differences between the maximum and the minimum values, no more big bumps all over the curve. This respects the traditional norms and decisions made upon the application of time of use pricing strategies (TOU) in case of non renewable energy production, when the provider usually adjusts his generation capacities to meet the purchasers' demand and seeks an almost constant level of demand. It is indeed the case, in this scenario S is similar to a traditional producer since he's not providing unconditional quantities all day long based on the desire of G. S will always wish to stick to what he has in store and keep the curve within a specific pattern.

Next in Game 1, Fig. 3a gives an idea about how the prices must be set to keep up with users' demand. And Fig. 4a shows how users react to that price. We see that first, when the price is too high, users lower their loads, therefore their actual load curve stays below the nominal curve. Second when electricity is cheap, they increase their loads, that is why the curve goes above the nominal values. Finally, when the price is just right, the two curves intersect allowing the nominal and the actual values to be equal. It is important to realize that at one point users' actual load curve doesn't exactly achieve the goals of Time of Use pricing because the curve isn't flattened as in the previous Fig. 4b. It is not flat nor follow a specific pattern. In fact, if this was about non-renewable resources it would be completely wrong. However, bearing in mind that in case of renewable resources (RE), the utility company's main purpose is to have users use up all the daily generation. Even if the load curve stays bumpy all day long, it won't matter as long as it copies the generation pattern.

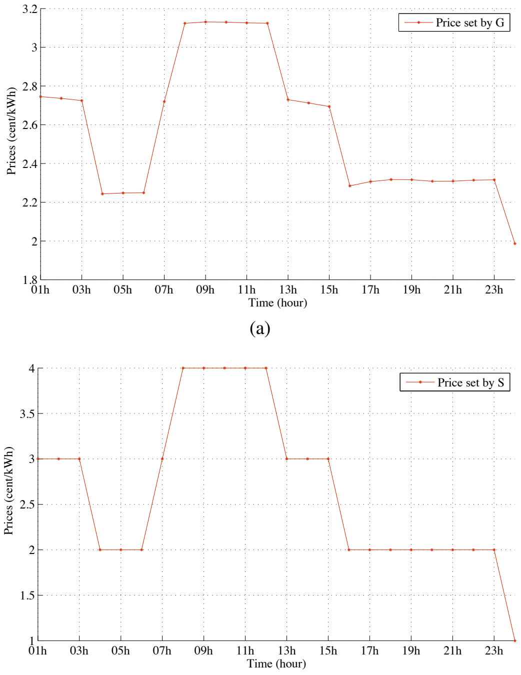
To conclude, Fig.5 gives us an idea on how all three curves ought to look like. It demonstrates how users' actual load curve is duplicating the wind curve and that the generation company's load curve is flattened. All of this has been done throught Time Of Use pricing strategies of both S and G.

V. Conclusion

In this paper, we have demonstrated how to achieve an optimal time of use pricing strategy of renewable energy (RE) with externalized storage and based on a model of three player games. We first defined and customized each players' objective function that gets him a better payoff. Next, we proposed a two games three players situation where the utility company G acts as a supplier of renewable energy (RE) towards users in the first game and as a mere user to the storage company S in the second game. Therefore, we constructed two identical games with the same objectives but different parameters. The results of our simulations supports our suggested idea. In fact, in game one, users' reactions to the prices of G copies the pattern of the generation curve, which means that they almost consume it all, while not having to pay much since the load curve itself is a result of an optimization problem. As for Game 2, the generation company's load curve that corresponds to the difference between their supply and users' demand is flattened and stayed within a constant level.

References

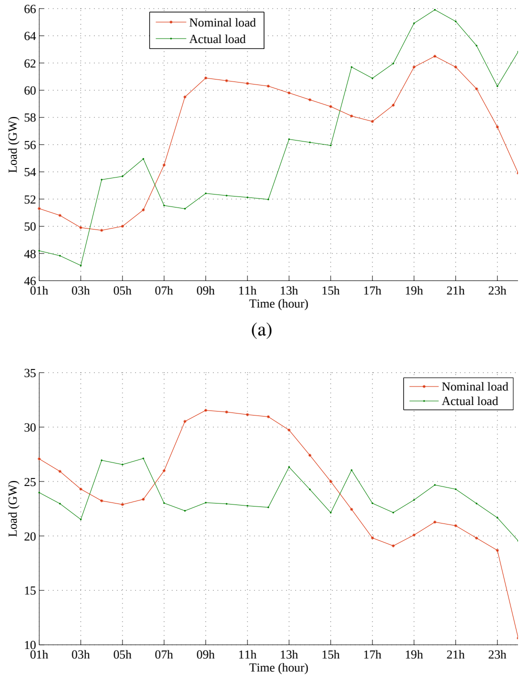
[1] C. Gellings, "The concept of demand-side management for electric utilities," *Proceedings of the IEEE*, vol. 73, no. 10, pp. 1468–1470, Oct. 1985.



(a)

(b)

Fig. 3: TOU pricing per kWh for both G (a) and S (b)



(a)

(b)

Fig. 4: Comparing the responses of U (a) and G (b) towards the TOU strategies with the nominal values

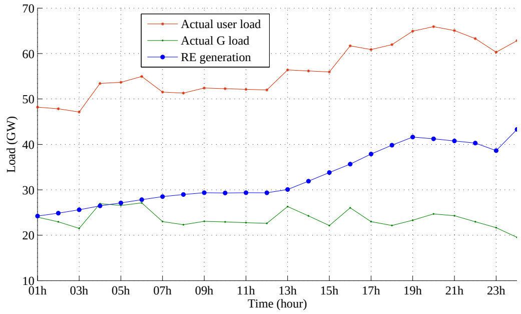


Fig. 5: Comparing the load curves of G and U and the renewable energy (RE) generation curve

[2] P. Palensky and D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[3] S. Zeng, Y. Ren, and J. Li, "A Game Model of Time-of-Use Electricity Pricing and its Simulation," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE, 2007, pp. 5045–5049.

[4] J. Li, "Research of time-of-use electricity pricing models in China: A survey," in *2008 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, pp. 2191–2195.

[5] I. Atzeni, L. G. Ordonez, G. Scutari, D. P. Palomar, and J. R. Fonollosa, "Demand-Side Management via Distributed Energy Generation and Storage Optimization," *IEEE Transactions on Smart Grid*, pp. 1–11, 2012.

[6] P. Yang, G. Tang, and A. Nehorai, "A game-theoretic approach for optimal time-of-use electricity pricing," *IEEE Transactions on Power Systems*, vol. 28, pp. 884–892, 2013.

[7] A. Haurie, R. Loulou, and G. Savard, "A two-player game model of power cogeneration in New England," *IEEE Transactions on Automatic Control*, no. 9, pp. 1451–1456.

[8] A. S. A. Awad, J. D. Fuller, T. H. M. EL-Fouly, and M. M. A. Salama, *IEEE Transactions on Sustainable Energy*.

[9] Admin, "Game Theory–Fudenberg.pdf," pp. 1–579.

[10] "Données de production éolienne - Elia." [Online]. Available: http://www.elia.be/fr/grid-data/production/production-eolienne

[11] "FERC: Electric Power Markets - MISO Daily Report Archives." [Online]. Available: http://www.ferc.gov/market-oversight/mkt-electric/midwest/miso-archives.asp

[12] "NREL: Energy Analysis - Energy Technology Cost and Performance Data." [Online]. Available: http://www.nrel.gov/analysis/tech_lcoe_re_cost_est.html

[13] K. E. Case and R. C. Fair, *Principles of economics*, 8th ed. Upper Saddle River, NJ : Prentice Hall, 2007, published in 2006.

[14] D. W. Caves and L. R. Christensen, "Econometric analysis of residential time-of-use electricity pricing experiments," *Journal of Econometrics*, vol. 14, no. 3, pp. 287–306, 1980.

[15] J. A. Espey, M. Espey *et al.*, "Turning on the lights: a meta-analysis of residential electricity demand elasticities," *Journal of Agricultural and Applied Economics*, vol. 36, no. 1, pp. 65–82, 2004.

[16] P. Poonpun and W. Jewell, "Analysis of the cost per kWh to store electricity," *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, vol. 23, no. 2, pp. 529–534, 2008.

[17] "Energy storage for variable renewable energy resource integration - A regional assessment for the Northwest Power Pool (NWPP)," *2011 IEEE/PES Power Systems Conference and Exposition, PSCE 2011*, 2011.

[18] V. Viswanathan, P. Balducci, and C. Jin, "National Assessment of Energy Storage for Grid Balancing and Arbitrage Phase II Volume 2: Cost and Performance Characterization," *Pnnl*, vol. 2, no. September, 2013.

WAVELENGTH CALIBRATION METHOD FOR SPECTRORADIOMETERS WITH PICOMETER UNCERTAINTIES

Zong, Y.
National Institute of Standards and Technology, Gaithersburg, Maryland, USA
yuqin.zong@nist.gov

DOI 10.25039/x44.2017.PP40

Abstract

Accurate wavelength calibration of array spectroradiometers is critical to many applications. We developed a new method for calibration of array spectroradiometers' wavelength scales using a wavelength-tunable kilohertz-pulsed optical parametric oscillator (OPO) laser and achieved a wavelength uncertainty on the level of picometers; a reduction of approximately two orders of magnitude compared to those using traditional methods. This high-accuracy wavelength calibration method can also be used to determine a spectroradiometer's pixel-to-pixel wavelength interval with an uncertainty of a few picometers, which is the key to achieving small uncertainties when a detector-based method is used for spectral calibration of array spectroradiometers using tunable lasers. Using this calibration method can also significantly reduce the overall measurement uncertainties in various applications.

Keywords: Wavelength Calibration, Pixel-to-Pixel Wavelength Interval, Spectroradiometer, Spectrometer, Spectrophotometer.

1 Introduction

Accurate wavelength calibration of array spectroradiometers is critical to many applications. For example, when measuring optical radiation of a light source using a spectroradiometer, a small wavelength error can result in a large spectral measurement error in the spectral region where the spectrum of the light source (e.g. a colour LED source) rises or falls sharply, or where the spectral responsivity of the spectroradiometer changes rapidly. An array spectroradiometer is typically calibrated for its wavelength scale by measuring a limited number (e.g. 10) of narrow spectral emission lines with known wavelengths and then determining the wavelengths for the corresponding pixels near the calibrated wavelengths. The wavelengths of the rest of the pixels, a majority, are obtained through interpolations, extrapolations, or curve-fitting based on the calibrated pixels. Using this conventional calibration approach, the wavelength uncertainty across the entire spectral range is limited from less than a nanometer (for metrology-grade spectroradiometers) to a few nanometers (for low cost spectroradiometers). In order to reduce the wavelength uncertainty, a much larger number of spectral lines is required for the wavelength calibration. Therefore, Fabry–Perot etalons and Lyot filters were proposed to be used for wavelength calibrations [Blattner et al. 2014] [Perret et al. 2010], both of which produce multiple transmission maxima over the spectroradiometer's spectral range. However, such devices are not readily available. Also, their transmission maxima are typically broad, which limits the wavelength calibration uncertainty.

2 New method for wavelength calibration

In this paper, we describe a new approach for calibration of spectroradiometers' wavelength scales that uses a fully-automated, kilohertz-pulsed optical parametric oscillator (OPO) laser. The OPO laser is tunable for wavelength from 210 nm to 2400 nm with a narrow bandwidth (*e.g.*, 0.08 nm at 350 nm, 0.14 nm at 600 nm, and 0.48 nm at 1100 nm), and it has been used at NIST for correction of stray light of spectroradiometers [Zong et al. 2006], and the calibration of detectors and spectroradiometers [Zong et al. 2012] [Zong et al. 2014]. The setup for the wavelength calibration is shown in Figure 1. A metrology-grade 1024-pixel CCD-array spectroradiometer with a fiber optic irradiance probe was calibrated for its wavelength scale using the OPO laser system. The spectral range of the spectroradiometer is from 300 nm to 1100 nm with a bandpass of approximately 2.5 nm. The OPO laser was tuned

Zong, Yuqin. "Wavelength calibration method for spectroradiometers with picometer uncertainties." Presented at CIE 2017 Midterm Meeting, Jeju, KR. October 23, 2017 - October 25, 2017.

across the entire spectral range of the spectroradiometer with a wavelength step of 5 nm. The wavelength of the OPO laser was measured by both the spectroradiometer (with specified wavelength accuracy of ±0.3 nm) and a high-accuracy laser spectrum analyzer (with specified wavelength accuracy of 3 pm at 350 nm wavelength, 7 pm at 600 nm, and 24 pm at 1100 nm). The spectral resolution ($\lambda/\Delta\lambda$) of the laser spectrum analyzer is 2×10^4, corresponding to a bandpass of 0.0175 nm at 350 nm, 0.03 nm at 600 nm, or 0.055 nm at 1100 nm which is much narrower than the bandwidth of the OPO laser. The software provided by the spectroradiometer’s manufacturer and that provided by the laser spectrum analyzer’s manufacturer were used for this calibration. The measured values from both instruments were peak wavelengths. The total number of measured laser lines was approximately 160. Because the wavelength calibration was fully automated, the total measurement time was less than one hour.

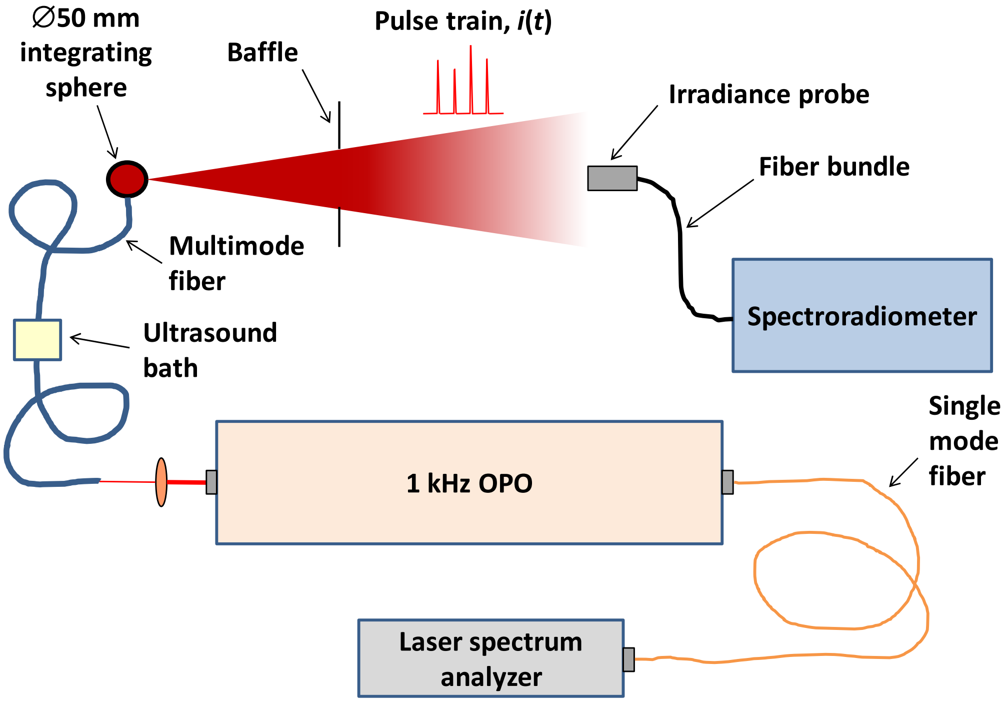


Figure 1 – Schematic for the wavelength calibration

3 Results of the wavelength calibration

The wavelength of the OPO laser measured by the spectroradiometer and that measured by the laser spectrum analyzer were compared. Using the laser spectrum analyzer as the reference standard, the wavelength error of the metrology-grade spectroradiometer at each measured wavelength was obtained. Figure 2 shows the wavelength errors of the spectroradiometer across its entire spectral range with the wavelength interval of 5 nm. The wavelength error varies from approximately -0.2 nm to 1 nm across the spectral range and changes rapidly in some spectral regions, indicating a fine-step wavelength calibration is required. The significantly larger than specified wavelength errors may be due to the replacement of the original fiber optic irradiance probe with a new one of the same type for this calibration. Using the determined wavelength errors at every 5 nm interval and the wavelength values of all pixels provided by the manufacturer based on the factory’s wavelength calibration, the wavelength error of each pixel was obtained by interpolation between measured wavelengths, and a correction for the wavelength error was applied to each pixel. As a result, the wavelength uncertainty of the spectroradiometer is significantly reduced. Note the wavelength calibration results may change slightly when a mathematic

Zong, Yuqin. ”Wavelength calibration method for spectroradiometers with picometer uncertainties.” Presented at CIE 2017 Midterm Meeting, Jeju, KR. October 23, 2017 - October 25, 2017.

method other than the peak wavelength method (such as centroid wavelength method or center wavelength method) is used because of spectroradiometer's asymmetric bandpass functions and possible asymmetric spectra of the kilohertz-pulsed OPO laser.

By knowing the accurate wavelength of each pixel, the pixel-to-pixel wavelength interval can be obtained. Figure 3 shows the gradual change of the pixel-to-pixel wavelength interval of the array detector of the spectroradiometer, which is approximately a linear function of pixel index number of the array detector. The uncertainty of the pixel-to-pixel wavelength interval can further be reduced to a level of a few picometers by curve-fitting the measured values which effectively eliminates the random measurement noises. Note the obtained pixel-to-pixel wavelength intervals do not depend on the mathematic method used for the wavelength measurements because the measurement errors associated with the mathematic method cancel at the two adjacent measured wavelengths.

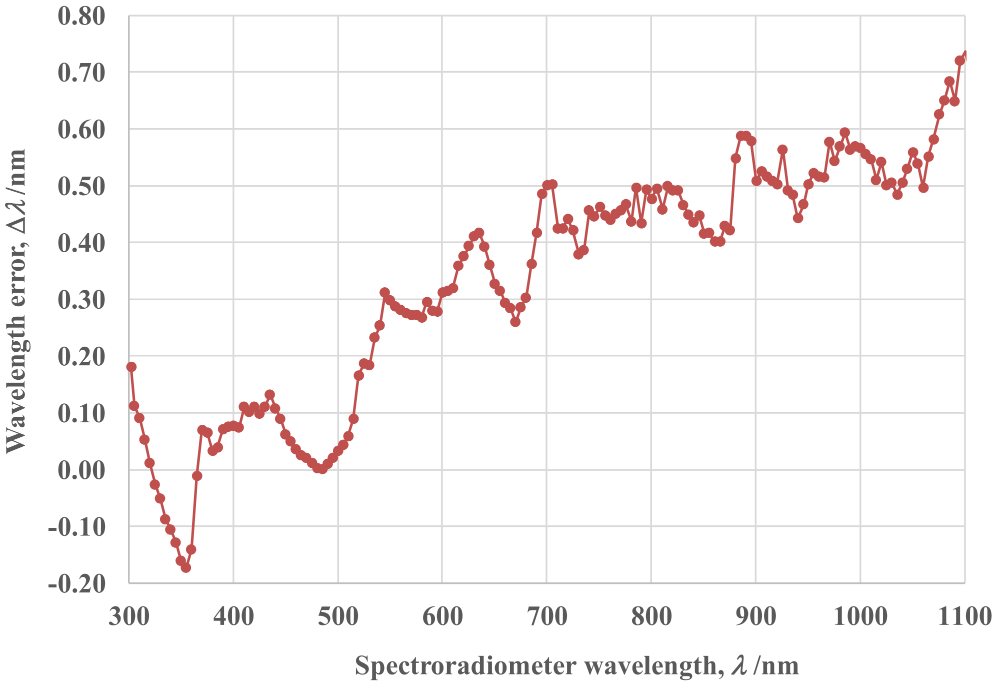


Figure 2 – Plot of the wavelength errors of a metrology-grade spectroradiometer

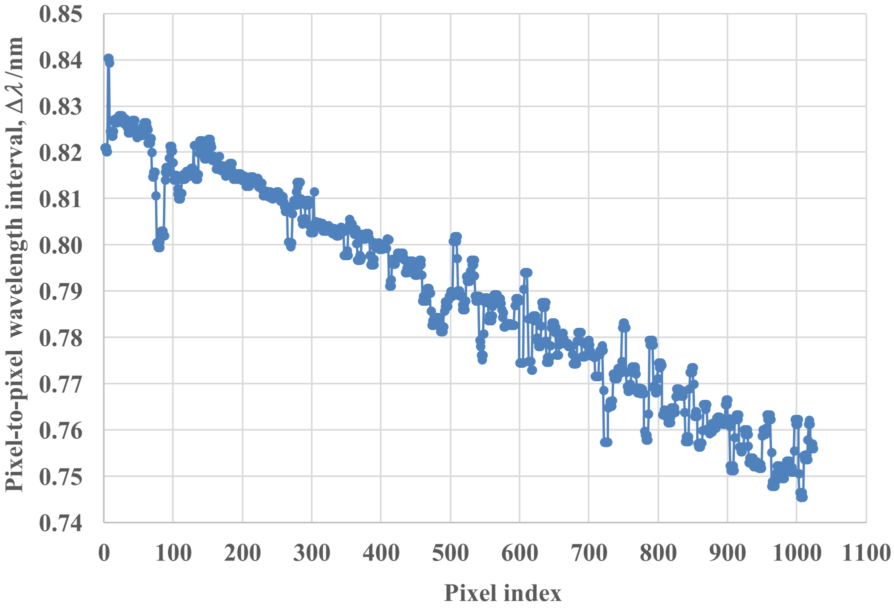


Figure 3 – Plot of pixel-to-pixel wavelength interval of the spectroradiometer

4 Summary

A new method for calibration of array spectroradiometers' wavelength scales was demonstrated using a fully-automated, kilohertz-pulsed, wavelength-tunable OPO laser. A calibration uncertainty of wavelength scale on the level of picometers can be achieved across the entire spectral range of the spectroradiometer; a reduction of approximately two orders of magnitude compared to the conventional approaches. This method can also be used to determine a spectroradiometer's pixel-to-pixel wavelength interval with an uncertainty of a few picometers, which is the key to achieving small uncertainties when the detector-based method is used for spectral calibration using the OPO laser [Zong et al. 2014]. High accuracy wavelength calibration of array spectroradiometers can also significantly reduce overall measurement uncertainties in many applications.

References

BLATTNER, P., FOALENG, S. M., VAN DEN BERG, S., GAWHARY, O. E., BLUMTHALER, M., GROBNER, J., and EGLI, L. 2014. Devices for characterizing the wavelength scale of UV spectrometers. Proceedings of NEWRAD 2014, S. Park, P. Karha, and E. Ikonen, eds., 201-202.

PERRET, E., BALMER, T., and HEUBERGER, M. 2010. Self-consistent algorithm for calibrating spectrometers to picometer accuracy over the entire wavelength range. Appl. Spectrosc., 64 (10), 1134-1144.

ZONG, Y., BROWN, S. W., JOHNSON, B. C., LYKKE, K. R., and OHNO, Y. 2006. Simple spectral stray light correction method for array spectroradiometers. Appl. Opt., 45 (6), 1111-1119.

ZONG, Y., BROWN, S. W., EPPELDAUER, G. P., LYKKE, K. R., and OHNO, Y. 2012. A new method for spectral irradiance and radiance responsivity calibrations using kilohertz pulsed tunable optical parametric oscillators. Metrologia, 49, S124-S129.

ZONG, Y., and SHAW P. 2014. Applications of pulsed OPO laser systems for optical radiometry. Proceedings of NEWRAD 2014, S. Park, P. Karha, and E. Ikonen, eds., 12-13.

Zong, Yuqin. "Wavelength calibration method for spectroradiometers with picometer uncertainties." Presented at CIE 2017 Midterm Meeting, Jeju, KR. October 23, 2017 - October 25, 2017.

Emerging Datasets and Analytics Opportunities in Metals Additive Manufacturing

Paul Witherell
National Institute of Standards and Technology
Gaithersburg, MD 20877 USA

Abstract

Additive manufacturing (AM) technologies continue to mature, evolving into stalwarts of high-end production lines, particularly with metals AM. Technology maturation has been facilitated by efforts in materials characterization, process sensing, and part qualification, among others. Advancements have been accompanied by a proliferation of AM data that is creating many new learning opportunities that have yet to be realized, hindered by a lack of curation and sharing. Data is often being generated in silos; associated with a specific time, process, material, location, etc. This manuscript investigates the state of data curation and analytics in AM. It begins by investigating AM data types and how this data is currently generated, curated, and shared. It then looks toward the future, where improvements in data curation will support emerging analytics. Finally, short-term needs and long-term opportunities are discussed, outlining future directions in data analytics for AM.

1 Introduction

In recent years, metals additive manufacturing (AM) technologies have established themselves as a stalwart of advanced manufacturing efforts. While AM technologies have long shown promise in realizing design freedoms, it has not been until recently that these technologies were seriously considered as production alternatives. Success stories from early adopters are demonstrating practicality and continuing to drive AM technologies into mainstream manufacturing as viable, profit-driven production technologies. Industry investments in metals AM have increased substantially in recent years and show no signs of stalling [4].

While initial investment has long been an entry barrier to metals AM, machine investment is only a portion of this cost. Much of the upfront costs can be attributed to “ramping up” activities, including the training and experience necessary to realize production-quality parts. The development of in-house expertise has been a distinguishing factor in determining maturity levels of AM practitioners. Improved understanding of the design-to-product transformation is necessary to overcome these barriers that hinder market entry and the capability to manufacture original and one-off designs.

Towards a better common understanding of AM processes, many efforts have focused on standards development, establishing specifications and communicating best practices in AM [5]. In parallel, significant investment has been made in measurements to reduce overall process uncertainty in AM, including design, material, process, and part characterization. Here we focus on the offshoots of these measurement, as AM characterization is introducing large amounts of new datasets from disparate sources. However, generating data and gaining knowledge from the data generated are quite separate matters.

Data analytics are critical to managing process variability and maturing AM processes. However, processing data can be a tedious task, and often becomes a limiting factor of the value of datasets produced. Studies have shown that only 30 percent of production data collected is also analyzed [6]. With new datasets being generated with domain-specific context, AM provides a unique opportunity to significantly increase that number. Recent advancements in software and informatics technologies are well positioned to capitalize on this proliferation of data.

This paper investigates new trends in AM data generation and methods to improve how AM data is leveraged. To increase data usage we must understand what the data sources are, what processing is necessary, and where the opportunities lie. Linking datasets across a build will provide new insight into the design, manufacture, and qualification details of an additively manufactured part. Compiling this data across builds promises to provide new insight into process control and process improvement.

2 Background

A signature trait of today’s advanced manufacturing processes is the ability to incorporate data into decision making. This incorporation includes measuring process performance, establishing performance baselines, integrating predictive analytics, defining performance metrics, and assessing quality, among others. AM, as perhaps the most “digital” of these advanced manufacturing technologies, stands to significantly benefit from advancements in how data is captured, curated, managed, and incorporated into decision making.

A single, unifying, characteristic of almost all AM data-driven activities is the desire to better understand

Witherell, Paul. ”Emerging Datasets and Analytics Opportunities in Metals Additive Manufacturing.” Presented at 4th Fraunhofer Direct Digital Manufacturing Conference (DDMC 2018), Berlin, DE. March 14, 2018 - March 15, 2018.

performance at each stage of an AM lifecycle [7]. Digital threads have been explored to establish provenance of part or process behaviors at given points of time, and their aggregates [8]. However, AM measurements are creating new data management challenges with increased levels of detail and new representation requirements of data types such as time series data and image registrations. The lifecycle stage-driven perception of a digital thread is being challenged as datasets increase in complexity.

New AM data requirements are often introduced from the bottom-up as new measurements are taken, diverging datasets across stages of the design-to-product transformation. Silos of disparate data become representative of ongoing data collection, with types of data varying from voltage signals to light intensities to images to statistical analyses to vectors to graphs to voxels. While the large amount of data clearly has much to offer, extracting value is the challenge now faced. What has yet to be established, but is emerging, is an emphasis on top-down configurations of these new sets of data.

Big data and related concepts are teaching us that as we collect new information there is always knowledge to be gained, even in the most unexpected places [9]. By emphasizing the learning potential of available datasets, new opportunities for gaining knowledge emerge. Significant hurdles must be overcome to effectively curate incoming AM data and fully exploit analytic opportunities. The next section discusses the datasets being generated across the lifecycle of an additively manufactured part, from raw material to final part.

3 Emerging Data

This section focuses on data collected from metals AM, particularly metal powder bed fusion. The datasets are separated into three subsections: feedstock material, in situ measurements, and ex situ measurements. This data may be collected through experiments, or in part production, to establish a greater empirical foundation.

3.1 Feedstock Material

The characterization of AM feedstock material has become increasingly advanced. Common powder measurements include powder size distribution and various flowability, morphology, and rheology measurements. New measurements and measurement techniques continue to emerge. Often described globally, powder size distribution and powder bed density are increasingly studied at various locations of the build volume [10]. Powder flow measurements, to study how powder spreads during the build, include the use of optical cameras mounted in a build chamber or on a recoater blade. Digital image correlation (DIC) measurements are capturing spread direction and powder velocity. Other powder spreading measurements include stiffness, coulomb damping, rolling friction, coefficient of restitution and angle of response during a powder sweep.

Characterization of powder particles includes chemical analysis (mass spectrometer) and morphology (X-ray Computed Tomography). Rheometer measurements have expanded to include metrics such as: total energy, permeability, torsion, normal force, and apparent density. Laser flash systems are measuring thermal properties of powder, including thermal diffusivity. Humidity and moisture measurements are studied during storage and processing.

Each of these measurements play a key part in characterizing feedstock powder, but their interrelationships, and how they affect the quality of a build, are not yet clear.

3.2 In Situ Measurements

Various in situ measurements are taken during the AM process, including those related to the material, process, and part. Measurements may come from both part production and experimental builds.

Material Layer- Layer-by-layer material monitoring provides insight into the state of feedstock material immediately before it is processed. In powder bed fusion, layerwise optical imaging using cameras above the build platform provides insight into the powder spread before processing. During spreading, measurements of the powder surface and of spreading angles are available with a profilometer. Layering instruments may also be monitored, such as the acceleration and vertical displacement of a recoater blade.

Melt Pool- Melt pool monitoring techniques are sought as a means for evaluating process parameters and providing insight into the final part during process time. Common melt pool measurements include: melt pool temperature, melt pool cooling rates, melt pool size, and melt pool shape.

High-speed thermal imaging cameras are used to measure light intensity for given spatial correlations, a simplification of temperature measurements (rather than true temperature). Supported measurements include melt pool temperatures, cooling rates, and generalizations of melt pool dimensions. Experimental studies (Figure 1) have investigated the effect of varying power, velocity, and scan strategy on melt pool dimensions and melt pool shape [1].

Photodetectors are used to detect light intensity from the build chamber (radiance over build volume) during the build and outputting voltage readings. Melt pools are being monitored with light-sensing cameras equivalent to a photodetector array. These cameras, coaxial with the laser path, are able to generate real time melt pool images and can provide data on melt pool shape, length, width, and location [11].

Measurements to help understand melt pool behavior are becoming increasingly advanced. Reflectance and emittance measurements of a melt pool are being measured with spectral direction emissivity methods. Emissivity, in combination with reflectance and radiance, has been measured to calculate true temperatures of a melt pool [12]. Emissivity values are specific to the AM process and measurement taken, with values dependent on material, measurement wavelength, melt pool temperature (energy density), direction (angle of observation) and surface characteristics (surface roughness).

Part- Part measurements taken in process are being directly correlated with final parts. Thermal measurements are providing insight into part cooling behavior, and optical measurements are being adopted for early defect detection.

3.3 Ex Situ Measurements

Measurements on a final part are critical for testing and qualification purposes. Additional measurements are often made on a witness specimen or witness coupons. These measurements may be destructive, as seen with mechanical testing, or non-destructive, as seen with scanning.

Surface Measurements- Surface measurements are used to qualify against specifications, correlate surface quality with process performance characteristics, and provide ex-situ characterization of melt pool surfaces and tracks, among other applications.

Outside of revisiting existing measurements, new surface characterizations are also being explored. Profile measurements, using techniques such as Scanning Electron Microscopy (SEM), allow detailed comparisons between modeled and produced parts. Datasets include surface images, surface height/ profile measurements, and information about other surface characteristics such as cracks and partially melted particles [13].

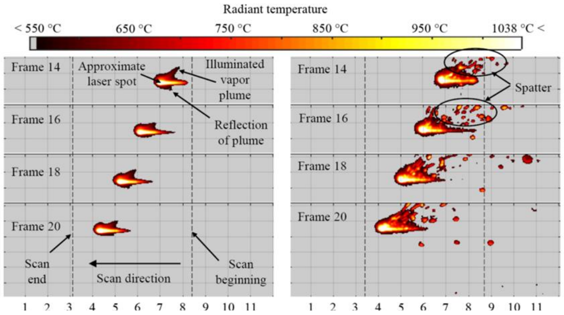


Figure 1: Time-series melt pool data from [1].

Microstructure Measurements- Ex situ measurements of part microstructure provide data for correlation with material and process measurements and can provide insight into part performance. These measurements, using methods such as SEM (Figure 2), support the evaluation of microstructure shape including grain orientation, grain size, and grain morphology of sample specimens.

Electron backscatter diffraction (EBSD) and neutron diffraction techniques are used to measure lattice spacing in atomic structures. The shape of a stressed unit geometry can be used to calculate stress and strain tensors from which residual stresses can be derived. Phase-specific stress/ strain measurements provide insight into how changes in crystalline phases affect part properties. The study of phase changes provides insight into the effectiveness of post processes used for stress relief, including heat treatment-induced phase transitions.

X-ray Computed Tomography (CT) measurements- X-ray CT imaging measurements are in the form of grayscale, voxel-based images (voxels located through coordinates) that can be translated to other formats such as STL. X-ray CT imaging allows for measuring

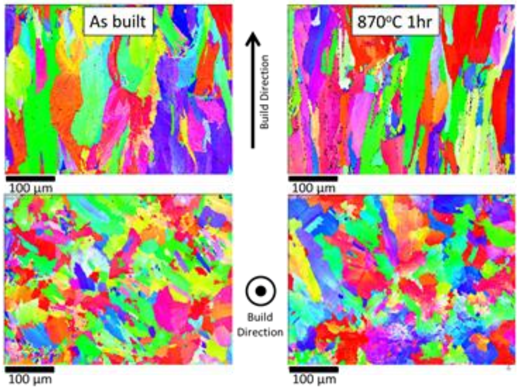


Figure 2: Microstructure images from [2].

Witherell, Paul. "Emerging Datasets and Analytics Opportunities in Metals Additive Manufacturing." Presented at 4th Fraunhofer Direct Digital Manufacturing Conference (DDMC 2018), Berlin, DE. March 14, 2018 - March 15, 2018.

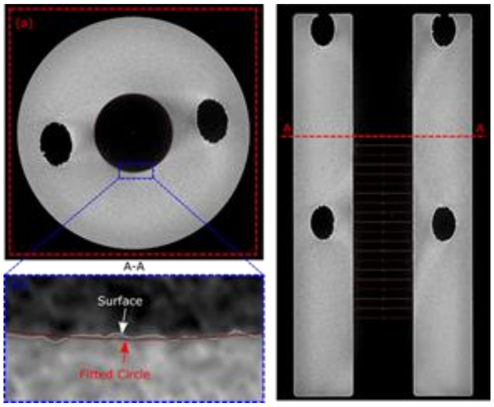


Figure 3: X-ray CT images from [3].

internal surfaces and volumes. Obtained images will vary based on material and scan parameters (e.g. voltage, line, magnification, face).

X-ray CT images (Figure 3) are used to identify and characterize part defects, including voids, porosity, features, and microcracks. Some characterizations are more challenging than others, such as with a “spider-web” like porosity with no well-defined shapes.

X-ray CT allows for the study of failure due to defect propagation initiated by voids. Mechanical testing (e.g. axial loading) of defect-induced specimens can be performed within an X-ray CT to provide insight into how pockets/shapes fail, essentially resulting in time series X-ray CT imaging.

3.4 Analytics from the Bottom-Up

Bottom-up measurements are often taken with specific requirements and analytics tasks. These tasks range from assessing equipment performance to assessing process behaviors to establishing fundamental AM correlations.

Model validation is a common application of measurements driven by bottom-up approaches. Thermal measurements are used to validate melt pool models. DIC measurements can validate powder spreading models. EBSD measurements help predict geometric errors and distortion due to residual stresses while validating distortion and stress models.

Often with multiple ways to measure a specific type of process behavior, measurements are sometimes desired to have insight into how different measurement types, or equipment types, compare. For instance, comparison of a thermal camera signature to a photodetector signature may be desired to see how well the measurements correlate.

When comparing in situ measurements with ex situ measurements, analytics provide insight into process controllability. For instance, in investigating scan strategies, correlations can be made between in situ (size, shape, temperature) and ex situ melt pool measurements (chevron, depth, microstructure, porosity). Surface measurements are being mapped to process characteristics, such as laser power and scan path. Ex situ track images provide information on scan patterns, track paths, cooling, texture, and melt pool profiles. Powder rheology measurements are correlated with powder spreading behaviors. Each of these examples provide insight into process control.

Perhaps the most sought after analytic opportunity, and one that requires both bottom-up and top-down approaches, is establishing expanded correlations between materials, process, part, and geometry. For instance, mechanical-microstructure property relationships are studied to identify correlations with part performance. Correlations between thermal history and microstructure lead to better understanding of how changes in microstructure occur.

While analytics already play an essential role in understanding AM processes, we are far from realizing the full learning potential of the data being collected.

4 Making Sense of it all: The Top-Down Approach

The requirement of a more top-down approach to AM data curation and analytics is based on the premise that to effectively utilize the enormously populous, yet greatly disparate, datasets, actions must be taken prior to the onset of data creation and curation. The goal is to ensure that data is curated in a manner that enables actionable, homogenously characterized data types to support increased analytic opportunities.

Reconciling disparate data types can be a daunting task. When working with many unknowns, asserting common references is problematic. The top-down approach offered here seeks to provide baseline references driven by domain nomenclatures, the domain being additive manufacturing. By initially constraining datasets, though abstractly, we can begin to label [14] the types of data being generated from the top-down. This approach seeks to quickly assert nomenclatures and associativity with datasets as the process dictates. Support for data characterization is perhaps best provided by advanced informatics techniques and algorithms, such as those associated with machine learning and neural networking, and enhanced representations, such as those provided through semantics and category theory.

Using the concepts just discussed, a top-down methodology for curating and analyzing AM data is proposed (Figure 4).

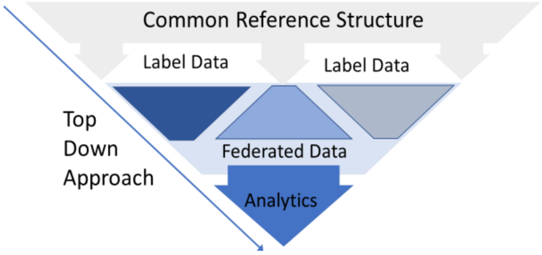


Figure 4: A top-down framework for AM data analytics.

Establishing a Common Reference Structure- A major challenge in effectively analyzing AM data is first establishing common references for how data is stored and accessed. With the data types varying significantly, challenges may come from software capabilities and data structure. The homogenization of heterogenous datasets requires some commonly shared identification mechanisms so that data can be effectively identified and labeled. A key to finding any common reference is utilizing structures that provide domain context and allow for controlled data label convergence.

One method for creating such structures is the adoption of ontology, where concepts can be explicitly defined and related. The explicit formalization of ontology supports well-structured definitions, and the contextualization of data, although the characterization of the data types remains a challenge. The hierarchical nature of the ontology supports the abstraction and subsumption of data labels, and thus the reconciliation of data types through labels.

Labeling and Registration- The post processing of raw data often necessitates human intervention and the methodical interpretation of datasets, a time-consuming process. Here we refer to labeling [14] as a mechanism for augmenting raw data with specific attributes or characteristics. The more attributes assigned to a piece of data or dataset, the better characterized the data becomes. Well-characterized data is essential for data analytics, as the characteristics provide the common context on which an analysis can be performed. Without this context, the meaning of the data can become lost, thus making analytics ineffective.

The time commitment of augmenting raw data can significantly limit the availability of data, especially time-dependent data. Recent advancements in computational techniques such as machine learning promise to lessen our reliance on human interpretation. Automation in the labeling process will effectively increase analytics opportunities.

One step beyond labeling is the process of data registration. Data registration allows data, such as images, to be linked through a single coordinate system. Figure 5 depicts what such a concept could mean if comprehensively applied to AM data. The datasets are registered across time (t), from model, to build data, to layer data, to process data, to microstructure data to X-ray CT data as a means to trace the evolution of a single part. This example of exhaustive traceability can lead to a better understanding of how defects in parts are formed.

Data Federation- The centralization of large datasets can quickly become unmanageable. Additionally, disparities in data types sometimes necessitate alternative storage methods. Utilizing data federation techniques is a necessary step to realizing the learning potential offered by AM datasets. Data federation [15] allows data stored in various locations to be accessed as a sole data source, a notion critical for big data analytics.

Given the variability demonstrated by AM processes, collaborative efforts are often sought to reduce individual investments. The concept of data federation facilitates collaboration, as various entities may share dispersed information and localized expertise.

Analytics- While AM data curation has long been a topic of interest, the advanced levels discussed here are not easily realized. However, they are necessary steps to take advantage of advanced analytics techniques that are becoming increasingly available. The larger and more diverse the datasets, the greater the potential to learn from them. While analytics opportunities exist within the objective-driven, bottom-up approach, greater opportunities are available. A top-down approach to data curation and analysis enables more objective learning approaches, where patterns may initially be sought irrespective of labels and characterization. Deep learning and neural networking techniques support such approaches, and properly preparing for such approaches is essential to realizing the upcoming opportunities.

Each of the steps described in Figure 4 are attainable with today’s tools. While much upfront investment is required to set the stage for new analytics opportunities, the payoffs are significant. Training methods with the ability to label and classify datasets will provide the instantaneous enrichment of new datasets. Data patterns will provide new insight into AM correlations and material-process-part relationships. Experimental and physics-based data can become complementary, feeding into new process control opportunities.

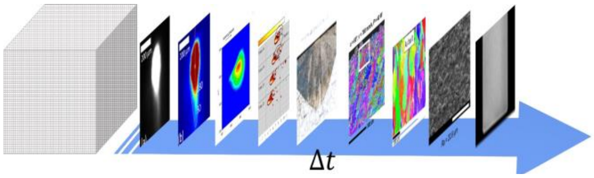


Figure 5. Data registration across time t.

5 Discussion

Until now, much of the contents of this paper have focused on emerging datasets and upcoming analytics opportunities often associated with big data. Big data techniques have demonstrated to be broadly applicable and extremely informative when correctly applied. Such approaches have the potential to revolutionize how data is used in manufacturing. However, most of the techniques discussed are susceptible to failure when data is mischaracterized or mislabeled. Methods are needed to guide, govern, and control the application of advanced analytics. Methods are also need to control data feeds, or to separate the collection of useful data from what may be considered valueless.

Closing the loop on the framework described in Section 4 requires determining and restricting behaviors at multiple scales. Consistently monitoring the conformance of datasets within the set constraints is important. The methods described can be applied in both sequence and in parallel, where errors can easily propagate and can become difficult to identify. Such outcomes restrict real time applications of data analytics. Desired scenarios support the generation and curation of data from the bottom-up while ensuring conformance from the top-down. Ideally a metalanguage provides the constructs with the ability to check for conformance across each of the steps described in Figure 4.

Metalanguages such as category theory can be applied to formally constrain interpretations of information, including those represented as disparate data and datasets. As a metalanguage based on mathematical formalisms, category theory has the ability to mathematically restrict the data and information flow outlined in Figure 4, thus constraining and controlling these transitions. Establishing such trust is crucial to taking the next steps in AM data analytics, where automation could be leveraged in applications such as inline control and programmed material behavior.

6 Summary

Additive manufacturing technologies have made significant advancements over the past five years. No longer viewed as only for toys or prototyping, significant investments have been made into maturing the technologies. One of the byproducts of these investments has been the enormous amount of data generated. This data will play a crucial role as AM technologies continue to mature. This paper investigated the characteristics of this emerging data and proposed a top-down methodology for curating this data in a way that will open new data analytics opportunities in the future.

Literature

[1] J. C. Heigel and B. M. Lane, "The effect of powder on cooling rate and melt pool length measurements using in situ thermographic techniques," in *Solid Freeform Fabrication Symposium*, 2017.

[2] T. Q. Phan, L. E. Levine, M. Stoudt, and J. C. Heigel, "Synchrotron X-ray Characterization of Powder-bed Fusion Laser Melt Traces on Solid Nickel-based Super Alloy Plates," in *TMS2017*, San Diego, CA, 2017.

[3] F. H. Kim, E. J. Garboczi, S. P. Moylan, and J. Slotwinski, "Investtigation of Pore Structure and Defects of Metal Addtive Manufacturing Components using X-ray Computed Tomography," *Additive Manufacturing*, 2017.

[4] T. Wohlers, *Wohlers report 2016*. Wohlers Associates, Inc, 2016.

[5] ANSI, "AMSC Standardization Roadmap for Additive Manufacturing," 2017, Available: https://www.ansi.org/standards_activities/standards_boards_panels/amsc/.

[6] S. Jacobson and R. Franzosa, "Predicts 2016: Opportunities Abound for the Factory of the Future to Reach Its Potential," vol. 25 November 2015, Available: https://www.gartner.com/doc/3172033/predicts--opportunities-abound-factory

[7] D. B. Kim, P. Witherell, R. Lipman, and S. C. Feng, "Streamlining the additive manufacturing digital spectrum: A systems approach," *Additive Manufacturing*, vol. 5, pp. 20-30, 2015.

[8] P. W. Witherell, Y. Lu, S. C. Feng, and D. Kim, "Towards a Digital Thread and Data Package for Metals Additive Manufacturing," *Journal of ASTM International*, 2017.

[9] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032-2033, 2012.

[10] J. Whiting and J. Fox, "Characterization of Feedstock in the Powder Bed Fusion Process: Sources of Variation in Particle Size Distribution and the Factors that Influence them," in *Solid Freeform Fabrication Symposium*, 2016.

[11] J. C. Fox, B. M. Lane, and H. Yeung, "Measurement of process dynamics through coaxially aligned high speed near-infrared imaging in laser powder bed fusion additive manufacturing," in *Thermosense: Thermal Infrared Applications XXXIX*, 2017, vol. 10214, p. 1021407: International Society for Optics and Photonics.

[12] B. Lane *et al.*, "Design, Developments, and Results from the NIST Additive Manufacturing Metrology Testbed (AMMT)," in *Proceedings of the 26th Annual International Solid Freeform Fabrication Symposium–An Additive Manufacturing Conference, Austin, TX*, 2016, pp. 1145-60.

[13] J. C. Fox, S. P. Moylan, and B. M. Lane, "Preliminary Study Toward Surface Texture as a Process Signature in Laser Powder Bed Fusion Additive Manufacturing," in *2016 Summer Topical Meeting: Dimensional Accuracy and Surface Finish in Additive Manufacturing*, 2016.

[14] H.-L. Chen, K.-T. Chuang, and M.-S. Chen, "On data labeling for clustering categorical data," *IEEE Transactions on knowledge and Data Engineering*, vol. 20, no. 11, pp. 1458-1472, 2008.

[15] J. Lee, P. Mazzoleni, J. Sairamesh, and M. Touma, "System and method for planning and generating queries for multi-dimensional analysis using domain models and data federation," ed: Google Patents, 2008.

The Industrial Ontologies Foundry (IOF) perspectives

Mohamed Hedi Karray[a], Neil Otte[b] Dimitris Kiritsis[c], Farhad Ameri[d], Boonserm Kulvatunyou[e], Chris Will[f], Rebeca Arista[g], Rahul Rai[h], and Barry Smith[i].

[a] *University of Toulouse, INP-ENIT, Tarbes, France*
[b] *Johns Hopkins University Applied Physics Laboratory, USA*
[c] *Ecole Federale Polytechnique de Lausanne, Switzerland*
[d] *State University of Texas, Austin, USA*
[e] *National Institute of Standards and Technology, Gaithersburg,USA*
[f] *Dassaults Systemes, Las Vegas, USA*
[g] *AIRBUS, Toulouse, France*
[h] *Clemson University, Clemson USA*
[i] *University at Buffalo, Buffalo, USA*

Abstract

In recent years there has been a number of promising technical and institutional developments regarding use of ontologies in industry. At the same time, however, most industrial ontology development work remains within the realm of academic research and is without significant uptake in commercial applications. In biomedicine, by contrast, ontologies have made significant inroads as valuable tools for achieving interoperability between data systems whose contents derive from widely heterogeneous sources. In this position paper, we present a set of principles learned from the successful Open Biomedical Ontologies (OBO) Foundry initiative to guide the design and development of the Industrial Ontologies Foundry (IOF), which is a counterpart to the OBO Foundry initiative for the manufacturing industry. We also illustrate the potential utility of these principles by sketching the conceptual design of a framework for sustainable IOF development.

Keywords 1

Ontology; industrial engineering; semantic interoperability; collaborative manufacturing; production engineering.

1. Introduction

Ontology has been touted as a solution to interoperability and a formal knowledge representation in an evolving collaborative industrial domain. However, even where ontologies are used in industry, they are often embedded as components in larger proprietary systems such that their existence remains unknown to the wider world. Moreover, such ontologies have been in almost all cases developed without any heed to issues of ontology reuse and to the lessons learned in earlier ontology initiatives [1]. Experiences in other fields such as bioinformatics or defense and intelligence, however, show that the most significant benefits of the ontology technology are derived from aggressive reuse of the same ontology content in multiple independent initiatives. The re-use of these developed ontologies from previous projects is not presently on the horizon, as most of the available ontologies we have discussed here are not interoperable, and classes across the ontologies are frequently redundant or used in multiple different ways. And yet, despite numerous benefits of ontologies, industrial companies have remained hesitant in using ontologies. Reconciling ontological researches between the industrial and academic domains is of utmost importance. Therefore, the focus of this work is on defining a framework for

Karray, Mohamed; Otte, Neil; Rai, Rahul; Ameri, Farhad; Kulvatunyou, Boonserm; Smith, Barry; Kiritsis, Dimitris; Will, Chris; Arista, Rebecca. "The Industrial Ontologies Foundry (IOF) perspectives." Presented at 10th International Conference on Interoperability for Enterprise Systems and Applications - Workshop5: Industrial Ontology Foundry (IOF) - Achieving Data Interoperability, Tarbes, FR. November 17, 2020 - November 19, 2020.

creating the IOF so that it may help in bridging the gulf that separates industry and academic ontology researches. In what follows, we outline a strategy to address this problem by drawing on the successful Open Biomedical Ontologies (OBO) Foundry initiative to guide the design and development of a foundry for industry: the Industrial Ontologies Foundry (IOF). We hypothesize that this strategy will lead, in an incremental fashion, to a comprehensive suite of interoperable software tools that will provide support for consistent data access and reasoning across the product life cycle.

2. Overview on ontologies and applications in industry

Ontology is a controlled vocabulary providing a consensus-based common set of terms for describing the types of entities in a given domain and the relations between them. It provides a common standardization platform, enabling information processing and exchange of data among both machines and humans.

High-quality domain ontologies are essential for their successful adoption in any domain. However, the field of ontology is nascent and therefore unstable. In practice, many new ontologists begin thinking as software developers who are accustomed to viewing things in terms of information and concepts, and not of things themselves and their natures. We argue that this orientation needs to change as we have also observed that such improper orientation has led to problems in many ontology development projects and has often led to a poor reputation for the ontology itself. Let us acknowledge that creating a new ontology and integrating existing ontologies are both time intensive tasks, and neither has well-documented and effective solutions [2]. Manually identifying relevant types of entities is always laborious and can be exasperating [3]. Extension or integration of existing ontologies is also quite difficult and costly [4]. Additionally, ontologies created by extending or integrating ontologies are often very weak [5]. There exist techniques and methodologies for automatic and semi-automatic ontology construction [6, 7]. However, these automated and semi-automated approaches are limited in scope. Most of the times, users are not fully satisfied with the results of the automated ontology learning techniques [8]. Manual work to validate and adapt the generated ontology is always needed. Another possible solution to facilitate ontology construction, is to reuse and integrate existing ontologies.

In manufacturing and industrial research, many collaborative projects involving industry and academia have been launched to provide ontology-based solutions to the problem of semantic interoperability across different industrial subfields. In a survey, we provide a table[2] which documents some of these projects, which involve both academy and industrial stakeholders. However, few industrial enterprises are adopting ontologies in their work environment, and most of the projects listed in the table are still at the level of research and have not been adopted as real-world solutions. Examples of collaborative efforts among industry and academia include OntoSTEP (based on ISO 10303), the Process Specification Language ontology (based on ISO 18629), and the Gas and Oil ontology (based on ISO 15926), and ONTO-PDM (based on ISO 10303 and IEC 62264). Relevant ontology content has been assembled also within the scope of Reference Architecture Model for Industry 4.0 (RAMI 4.0) project, including the Standards Ontology (STO)[3] and the RAMI Vocabulary Ontology.[4] In none of these cases, however, there is no industrial reuse of the mentioned ontologies on the horizon. They are in almost all cases not interoperable with each other; the same terms are frequently used in different ontologies in different ways or different terms are used in the same way. Taking the example of the class product: this class is presented differently in different ontologies because they adopt different perspectives or conceptualizations (deriving from design, manufacturing, maintenance, sales, and so forth). Yet all of them are referring to more or less one and the same thing in reality. In PSL, a product is defined as: *An object that satisfies a design specification*. In the PRONTO Product ontology [9] it is defined as: *An abstraction representing individual items having physical existence*. In the Event Ontology as: *Everything produced by an event*. In schema.org as: *Any offered product or service* (with

[2] https://industrialontologies.org

[3] https://github.com/i40-Tools/StandardOntology

[4] http://i40.semantic-interoperability.org/rami/Documentation/index.html

examples: a pair of shoes; a concert ticket; the rental of a car; a haircut; or an episode of a TV show streamed online). And in Onto-PDM as equivalent to *MaterialDefinitionType* in IEC:62264.

3. Toward reusing and sharing of ontologies in industry

The goal is to create a suite of principles-based ontologies conforming to a hub-and-spokes model. The hub will contain a small number of reference ontologies that are non-redundant in the sense that they assert no terms in common. Connected to the hub in increasingly widening circles will be a much larger and constantly expanding number of application ontologies, all of which draw on the application ontologies closer to the hub, and ultimately on the reference ontologies at the center, in defining their terms. All the ontologies in the resulting suite will be interoperable, in a precisely defined sense, in virtue of their adherence to the common set of principles. The suite would then constitute The Industrial Ontologies Foundry (IOF)[5], a foundry for industry constructed along the lines of the OBO (Open Biomedical Ontologies) Foundry. The idea is that multiple parties agree to use one another's ontologies, to share a common set of principles, and to share the work of revising both ontologies and principles as these are tested in use with ever-expanding bodies of data, thus ensuring that the result achieves the required degree of interoperability. These benefits can be gained, however, only if the ontologies are developed in such a way as to form an open suite of ontology modules that have been developed in tandem in a way that ensures interoperation

3.1. The proposal of the Industrial Ontologies Foundry (IOF)

The IOF will provide a framework to focus collaboration efforts on developing, standardizing, sharing, maintaining, updating, and documenting industrial ontologies. The main aim of this foundry is to meet the needs of industrial stakeholders by providing a reliable turnkey solution and giving them best practices to integrate ontologies in their businesses.

In other words, such a solution will provide:

- Fully open source stable ontologies.
- Clear and well-documented ontologies.
- Scenarios in which industries will find it advantageous to reuse ontologies, terminologies, and coding systems that have been tried and tested.
- Prospective standardizations built with a coherent top-level ontology and with content contributed and monitored by domain specialists.

Each ontology in the foundry will be managed by a working group that will ensure the collaborative development and the integration of the ontology according to foundry principles. The working group will also edit the documentation about the ontology and define use cases and scenarios of its use. The working group will handle the update and the maintenance of the ontology. As we have mentioned, the objective of the foundry is to have one standard ontology for each domain, so any suggestion to modify or update the ontology must be communicated with the working group that discusses the utility of this update. The working group will be composed of domain experts and users of the adopted top-level ontology. The technical board will involve both senior ontologists and senior domain experts. The role of these boards is to apply the peer review process of the foundry to check the ontologies edited by the working groups according to the IOF principles. Working groups will be created according to a specific process application.

[5] https://www.industrialontologies.org

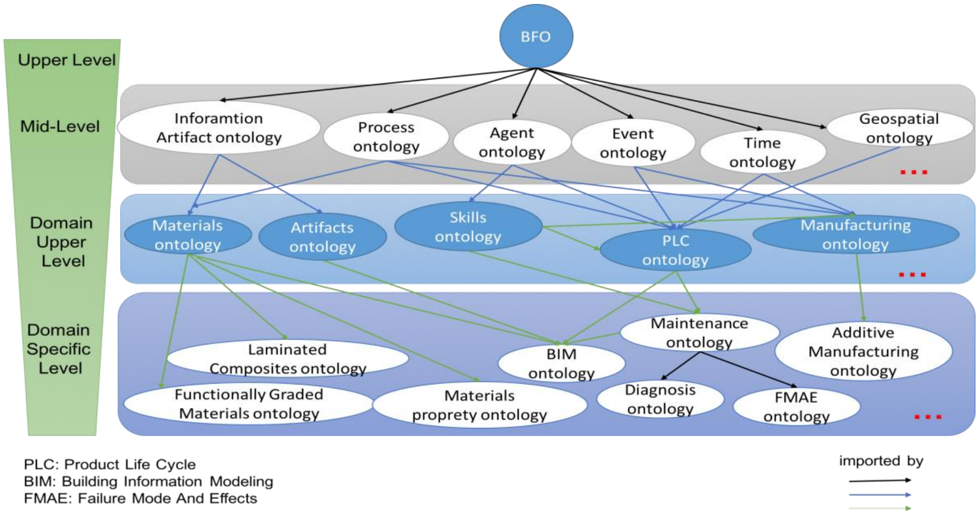


Figure 1: Classification of the proposed IOF core ontologies

The IOF ontologies may be organized according to four levels designated ‘Upper level,’ ‘mid-level,’ ‘domain-upper level,’ and ‘domain-specific level’ (see Figure 1). The aim of this separation is to ensure greater interoperability and reuse while allowing for the development of domain ontologies. As showed in [10], using such a suite of ontologies, professionals in industry can develop their customized application ontologies. These ontologies can be connected to and may reuse, one another. An ‘upper ontology’ or ‘top-level ontology’ is defined as a high-level, domain-independent ontology, providing a framework by which more domain specific ontologies may be derived [11]. These are also sometimes called ‘foundational ontologies,’ and can be compared to the meta-model of a conceptual schema [12]. A mid-level ontology provides more concrete representations of abstract entities defined in the upper-level ontology. It serves as a bridge between abstract entities defined in the upper-level ontology and the domain ontology [12]. A domain upper level ontology specifies classes particular to a domain of interest and represents those concepts and their relationships from a domain-specific perspective [12]. There may be two layers of domain ontology maintained by IOF as shown in Figure 1. The specialization of a domain ontology (domain specific level) is called an ‘application’ or ‘local ontology’, see Figure 2. This type of ontology represents a domain according to a single viewpoint of a user or a developer [13].

Basic Formal Ontology (BFO) [14] is the selected candidate upper-level ontology for the Industrial Ontologies Foundry. BFO is a small ontology, containing about 35 terms, whose role is primarily to work behind the scenes, imposing a perspective on the classes that extend from them. Instead, it is in the low-level ontologies that we will find those terms that predominate in practical uses of the ontology [15]. BFO is now approved as an ISO standard (ISO 21838-2).

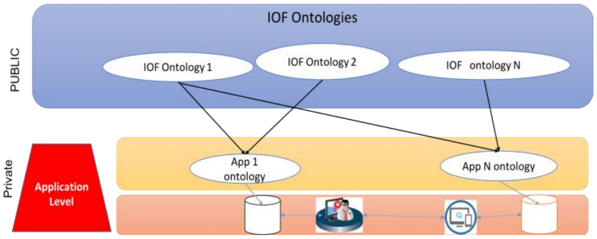


Figure 2: Application interoperability gained through reuse of IOF ontologies

From an operational point of view, as shown in figure 2, industrial users can adopt IOF ontologies to build their own network of ontologies by important terms (with definitions) from different ontologies in the Foundry. In some cases, they will import entire modules from the IOF registry, and then connect them to obtain an application ontology for their own business case. The IOF will then promote the interoperability of the modules in this network.

3.2. The IOF roadmap

To manage the scope and expectations, the IOF community kicked-off its effort with a proof-of-concept (POC) project [16]. This project was intended to test the feasibility of IOF goals. Therefore, the objectives of the POC included not only producing a small initial ontology, but also testing the organizational structure (described above) and producing and testing drafts principles and guidelines. To set the scope, the POC started by asking for most interested manufacturing-related terms from the community. After collecting all the submissions, 20 terms were identified based on the frequencies of matches across the submissions. Each term has a (synonym) set of closely matched terms; therefore, the output of this step is called the top-20 set. According to the objectives of POC, five Working Groups have been created. The Top-down WG, is responsible for providing consistent terms and definitions of high-level entities used across other WGs. Top-down WG started formalizing top-20 set polling from the first step of POC project. The Supply Chain (SC) WG is motivated by use cases such as supplier discovery (i.e., supplier capability matching with manufacturing requirements), supply material traceability. The Maintenance WG is motivated by a few use cases including the maintenance strategy assurance, asset operator failure mode and effects analysis (FMEA), and predictive maintenance [17]. The Process Planning and Production Scheduling (PPS) WG is motivated by use cases including process planning, manufacturability analysis, shopfloor design, and production scheduling. The Product Service System (PSS) working group (WG) aims to create a basis ontology for enhancing engineering of PSS in manufacturing, by modelling all the aspects that affect, or could affect a PSS. In addition to the weekly online meetings, the WGs use a set of team collaboration tools to share models, discuss terms and definitions. A technical principles document[6] has also been developed to guide the design compliance across WGs.

4. Conclusion

This work has proposed the IOF as a strategy for coordinating the development of ontologies, addressing interoperability in the industrial domain, and overcoming the reticence to rely on ontologies as reliable, turnkey solutions. This reticence is due to the problems that persist in ontology engineering such as building methodologies, reusability, integration, as well as costs and dependability. Present research is clear in concluding that existing ontologies suffer from a lack of interoperability. In almost all cases, these ontologies are developed independently, with no reuse of ontology work from the outside and no attempt to profit from lessons learned in earlier initiatives. Hence, they cannot be exploited as a reference in an industrial large scale. In this paper, we have presented a strategy to provide an open ontology framework, called the Industrial Ontologies Foundry (IOF), involving a suite of principles-based ontologies, which broadly represent a hub-and-spokes model. The proposed IOF will provide industrial stakeholders a reliable turnkey solution and give them the best practices to integrate ontologies in their businesses.

5. References

[1] S. Borgo, L. Lesmo. “The attractiveness of foundational ontologies in industry.” Frontiers in Artificial Intelligence and Applications. 174 (1). 2008.

[6] https://www.industrialontologies.org/technical-principles/

[2] T. Wächter, G. Fabian, M. Schroeder, "DOG4DAG: semi-automated ontology generation in obo-edit and protégé". 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences, ACM,: 119-120. December 2011.

[3] K. Xiangping, D. Li, S. Wang. "Research on domain ontology in different granulations based on concept lattice." Knowledge-Based Systems 27: 152-161. 2012.

[4] L. Obrst, , M. Grüninger, et al. "Semantic web and big data meets applied ontology." Applied Ontology. 9; 2; 8-; 155-170. 2014.

[5] L. Zhao, I. Ryutaro Ichise. "Ontology integration for linked data." Journal on Data Semantics 3, no. 4: 237-254. 2014.

[6] O. Medelyan, I.H. Witten, A. Divoli, J. Broekstra, "Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3 (4): 257-279. 2013.

[7] D. Küçük, Y. Arslan, "Semi-automatic construction of a domain ontology for wind energy using Wikipedia articles." Renewable Energy, 62, 484-489. 2014.

[8] S. Mittal, N. Mittal. "Tools for ontology building from texts: Analysis and improvement of the results of text2onto". IOSR Journal of Computer Engineering (IOSR-JCE), pp 2278–0661, 2013.

[9] M. Vegetti, H. Leone, G. Henning , "PRONTO: An ontology for comprehensive and consistent representation of product information." Engineering Applications of Artificial Intelligence, 24 (8), 1305-1327. 2011.

[10] C. Palmer, Z. Usman, O. Canciglieri Junior, A. Malucelli and R. I. M. Young. "Interoperable manufacturing knowledge systems". International Journal of Production Research, 56(8), 2733-2752. 2018.

[11] C. Phytila. "An Analysis of the SUMO and Description in Unified Modeling Language". HTM. April 2002

[12] R. Poli, M. Healy and A. Kameas. "Theory and applications of ontology". Computer applications, Dordrecht: Springer, pp. 1-26. 2010.

[13] C. Roussey, F. Pinet, M.A. Kang, O. Corcho. "An introduction to ontologies and ontology engineering." Ontologies in Urban Development Projects, pp. 9-38. 2011.

[14] R. Arp, B. Smith and A.D. Spear. "Building Ontologies with Basic Formal Ontology", MIT Press. 2015.

[15] B. Smith and M. Brochhausen. "Putting biomedical ontologies to work." Methods of information in medicine, 49 (2), 135. 140. 2010.

[16] B. Kulvatunyou, E.K. Wallace, D. Kiritsis, B. Smith, and C. Will. "The Industrial Ontologies Foundry Proof-of-Concept Project". International Conference Advances in Production Management Systems (APMS 2018). 2018.

[17] M. H. Karray, F. Ameri, M. Hodkiewicz, T. Louge. "ROMAIN: Towards a BFO compliant Reference Ontology for Industrial Maintenance". Applied Ontology, 14 (2) (2019): 155-157. 2019.

Dynamic measurement of nanoliter per minute flow by scaled dosage of fluorescent solutions

Gregory A. Cooksey, Paul N. Patrone, James R. Hands, Stephen Meek, Anthony Kearsley

We report a new method that permits continuous, in-situ flow measurements in a microchannel over the range of 10 nl/min to 10 µl/min. Notably, our approach requires little information about or control over the microchannel geometry, yet it achieves relative uncertainties at an unprecedented level of <20% throughout the entire range. Microfluidic devices are increasingly being utilized for real-time chemical and biological measurements on the microscale, thereby reaching new milestones in signal detection and experiment complexity. However, achieving reproducibility and determining measurement accuracy are substantial challenges in many applications of these technologies, including those intended for medical uses. Precise measurements of flow are fundamental to flow-based systems, but dynamic *in situ* measurement of flow remains a challenge below the microliter per minute regime (Ahrens et al. 2014). Previously we reported a device that measures volumetric displacement of a "stripe" of caged fluorescein between an uncaging region and a fluorescence measurement region within a microfluidic structure containing integrated optical waveguides (Hands et al. µTAS 2017). Though traceable to fluid volume (m^3) and time (s), the uncertainties of those measurements are limited by, for example, how well one can measure and uniformly control flow channel dimensions, optical power density and collection efficiencies from the optical waveguides. Because the system must operate in a convection-dominated regime, determination of flows <1 µl/min become impractical due to requirement of smaller channel dimensions, and thus expanding relative uncertainties compared to typical dimensional uncertainties > 1µm.

To realize nanoliter per minute flow measurements, we have simplified our previous design (Hands et al. 2017) to no longer require an activation light source (**Figure 1**) and create contrast in the flowing liquid by adding fluorescein, a small molecule that emits light with limited excitation. Critically, fluorescein is also destroyed in the presence of high excitation power, as demonstrated in **Figure 2**. In this context, we have identified a mathematical relationship that, under reasonable assumptions (e.g. Poiseuille flow), permits the use of scaling arguments to measure flow-rates. We demonstrate that the fluorescence efficiency (fluorescence emission/excitation power) is determined uniquely by the dosage of excitation light, i.e. excitation intensity times the residence time in the light path (inverse flow-rate) (**Figure 3**). Namely, as dosage increases, fluorescein molecules photobleach, resulting in a drop in fluorescence efficiency that determines the flow rate given a known excitation intensity. Surprisingly, our analysis indicates that this approach is accurate having only order-of-magnitude estimates of the device geometry and operating conditions, with only a single-point calibration being used to establish an absolute scale. Moreover, such methods have the interesting property of *decreasing* absolute uncertainties when measuring flow rates less than the calibration point. Experimental results confirm the validity of these results over 3 orders of magnitude (≈ 10 to 5000 nl/min) with less than 20% relative uncertainty. Future work involves modification of the flow system to enable use of the flowmeter in-series with microfluidic networks without contamination by fluorescein and to improve device robustness to stray light and to channel deformation at increased flow.

Cooksey, Gregory A.; Patrone, Paul; Hands, James; Meek, Stephen; Kearsley, Anthony J. "Dynamic Measurement of Nanoliter Per Minute Flow by Scaled Dosage of Fluorescent Solutions." Presented at 22nd International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTas 2018), Kaohsiung, TW. November 11, 2018 - November 15, 2018.

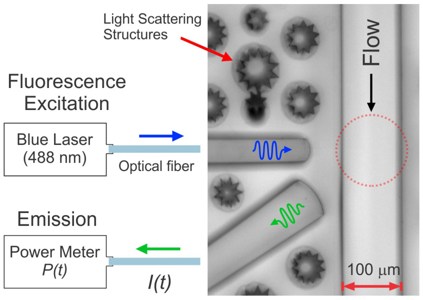


Figure 1. Schematic of optofluidic flow meter made of PDMS and containing integrated waveguides filled with optical adhesive (Norland #88). A narrow waveguide delivers excitation light to fluorescein in a flow channel while a wider waveguide collects emission light and couples to a power meter (Newport 2936-R) through an emission filter. Flow channel cross section is 100 µm x 80 µm.

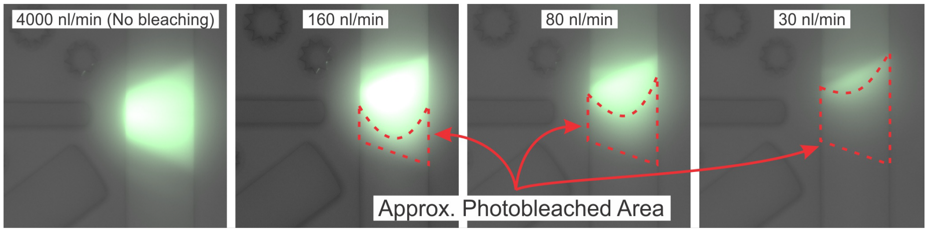


Figure 2. Microscopy images of different flow rates exposed to 488 nm laser light (15 mW). At high flow rates, fluorescein molecules move through the laser quickly and do not photobleach, thus showing the profile of the excitation light through the channel (LEFT). As flow rate decreases, light dosage increases and photobleaching becomes evident, particularly near the channel walls where flow velocity is slow.

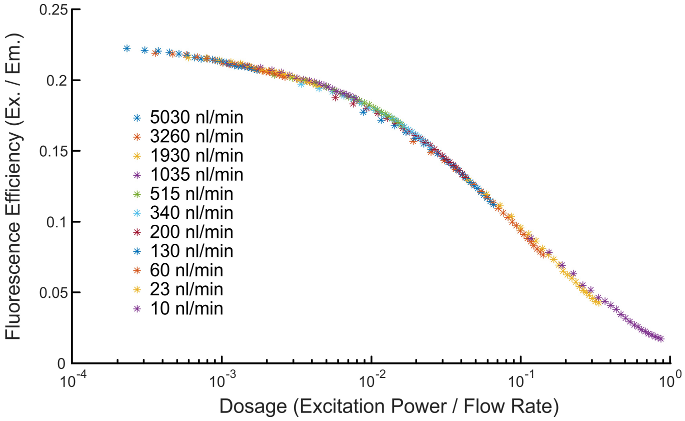


Figure 3. Fluorescence efficiency (emission per excitation power) was measured in 10% increments of laser power from 0 to 15 mW at various flow rates, which were recorded on a flowmeter with 10% uncertainty (Sensirion). The dosage calibration shows preservation of the scaling relationship covering nearly 4 orders of magnitude in a single device. Using this relationship, fluorescence emission can be continuously monitored at constant excitation intensity to determine absolute flow rate.

Cooksey, Gregory A.; Patrone, Paul; Hands, James; Meek, Stephen; Kearsley, Anthony J. "Dynamic Measurement of Nanoliter Per Minute Flow by Scaled Dosage of Fluorescent Solutions." Presented at 22nd International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTas 2018), Kaohsiung, TW. November 11, 2018 - November 15, 2018.

REFERENCES:

1) M. Ahrens, St. Klein, B. Nestler, C. Damiani. Measurement and Science Technology. 25, pp 1-9, 2014
2) Hands JR, Cooksey GA (2017) "Integrated Optical Waveguides for *in situ* Microflow Measurements." *Proc MicroTAS 2017*, 1513-1514

Cooksey, Gregory A.; Patrone, Paul; Hands, James; Meek, Stephen; Kearsley, Anthony J. "Dynamic Measurement of Nanoliter Per Minute Flow by Scaled Dosage of Fluorescent Solutions." Presented at 22nd International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTas 2018), Kaohsiung, TW. November 11, 2018 - November 15, 2018.

Neutron Spin Rotation Measurements

M. Sarsour[1,*], *J.* Amadio[2], *E.* Anderson[3], *L.* Barrón-Palos[4], *B.* Crawford[2], *C.* Crawford[5], *D.* Esposito[6], *W.* Fox[3], *I.* Francis[7], *J.* Fry[8], *H.* Gardiner[9], *C.* Haddock[10], *A.* Holly[11], *S.F.* Hoogerheide[12], *K.* Korsak[3], *J.* Lieers[13], *S.* Magers[2], *M.* Maldonado-Velázquez[4], *D.* Mayorov[14], *H.P.* Mumm[12], *J.S.* Nico[12], *T.* Okudaira[10], *C.* Paudel[1], *S.* Santra[15], *H.M.* Shimizu[10], *W.M.* Snow[3], *A.* Sprow[5], *K.* Steen[3], *H.E.* Swanson[16], *F.* Tôvesson[14], *J.* Vanderwerp[3], and *P.A.* Yergeau[2]

[1]Georgia State University, Atlanta, GA 30303, USA
[2]Gettysburg College, Gettysburg, PA 17325, USA
[3]Indiana University, Bloomington, IN 47408, USA
[4]Universidad Nacional Autónoma de México, D.F. 04510, México
[5]University of Kentucky, Lexington, KY 40506, USA
[6]University of Dayton, Dayton, OH 45469, USA
[7]612 S Mitchell St Bloomington, IN 47401, USA
[8]University of Virginia, Charlottesville, VA 22903, USA
[9]Louisiana State University, Baton Rouge, LA 70803, USA
[10]Nagoya University, Furocho, Chikusa Ward, Nagoya, Aichi Prefecture 464-0814, Japan
[11]Tennessee Tech University, Cookeville, TN 38505, USA
[12]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[13]Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
[14]Los Alamos National Lab, Los Alamos, NM 87545, USA
[15]Bhabha Atomic Research Centre, Trombay, Mumbai, Maharashtra 400085, India
[16]University of Washington, Seattle, WA 98105, USA

Abstract. The neutron spin rotation (NSR) collaboration used parity-violating spin rotation of transversely polarized neutrons transmitted through a 0.5 m liquid helium target to constrain weak coupling constants between nucleons. While consistent with theoretical expectation, the upper limit set by this measurement on the rotation angle is limited by statistical uncertainties. The NSR collaboration is preparing a new measurement to improve this statistically-limited result by about an order of magnitude. In addition to using the new high-flux NG-C beam at the NIST Center for Neutron Research, the apparatus was upgraded to take advantage of the larger-area and more divergent NG-C beam. Significant improvements are also being made to the cryogenic design. Details of these improvements and readiness of the upgraded apparatus are presented. We also comment on how recent theoretical work combining effective field theory techniques with the $1/N_c$ expansion of QCD along with previous NN weak measurements can be used to make a prediction for $d\phi/dz$ in ^{4}He.

An experiment using the same apparatus with a room-temperature target was carried out at LANSCE to place limits on parity-conserving rotations from possible fifth-force interactions to complement previous studies. We sought this interaction using a slow neutron polarimeter that passed transversely polarized slow neutrons by unpolarized slabs of material arranged so that this interaction would tilt the plane of polarization and develop a component along the neutron momentum. The results of this measurement and its impact on the neutron-matter coupling g_A^2 from such an interaction are presented.

1 Introduction

The nucleon-nucleon (NN) weak interaction is one of the most poorly understood areas of the Standard Model. Understanding the NN weak interaction in general and NN weak interaction amplitudes in particular is very important for several reasons [1–4]. It is a probe of Quantum Chromodynamics (QCD) that samples the short-distance quark-quark correlations in QCD and its chiral symmetry-dominated long-range properties. Knowledge of weak NN couplings can shed light on parity violation phenomena in atomic and nuclear physics [5–8]. The theoretical advances in the description of NN weak interaction set the stage for eventual quantitative predictions of the weak interaction directly from the Standard Model and QCD and thus make better contact with QCD in the near future.

The uncertainty in details of the hadronic weak interaction (HWI) stems largely from the very short range of the quark-quark weak interaction (~0.01 fm) compared with the range of nucleon-nucleon interaction (~1 fm) and the relative strength of the weak to strong force between strongly interacting particles (10^{-7} – 10^{-6}). Experimentally one uses parity violation to isolate the weak con-

[*]e-mail: msar@gsu.edu

Nico, Jeffrey S.; Hoogerheide, Shannon; Mumm, Hans Pieter; Sarsour, Murad; Amadio, J; Anderson, Eamon; Barron-Palos, Libertad; Crawford, Bret; Crawford, Chris; Esposito, D.; Fox, Walter; Francis, I; Fry, J; Haddock, Chris; Holley, Adam; Korsak, Kirill; Lieers, J; Magers, S; Maldonado-Velazquez, M.; Mayorov, D; Okudaira, T; Paudel, C; Santra, S; Shimizu, H.M.; Snow, William; Sprow, A.; Steen, K.; Swanson, H; Vanderwerp, John; Yergeau, P. "Neutron Spin Rotation Measurements." Presented at International Workshop on Particle Physics at Neutron Sources (PPNS 2018), Grenoble, FR. May 24, 2018 - May 26, 2018.

tribution. The primary analysis tool over the past few decades has been the meson exchange model of Desplanques, Donoghue, and Holstein (DDH) [9, 10]. The DDH models the HWI between nucleons as an exchange of a light meson (π, ρ, ω) where the meson couples to the nucleons via strong coupling at one vertex and weak coupling at the other. This model leads to seven meson coupling constants, h_π^1, $h_\omega^{0,1}$, $h_\rho^{0,1,1',2}$, which need to be fixed by measurement. Several measurements from NN system to heavy nuclei were carried out but only some place constraints on HWI [11], which include $\vec{p}$+p [12–17], $\vec{p}$+^{4}He [18, 19], P_γ(^{18}F) [20–23], and A_γ(^{19}F) [24–26]. In heavy nuclei measurements, an asymmetry enhancement is observed but the theory calculations are complicated [5, 27]. This difficulty motivated studying few body systems because while they are expected to have small asymmetries, 10^{-7} – 10^{-8}, the associated theory is calculable and reliable.

NPDGamma experiment is the first few nucleon precision measurement of h_π^1 with little or no theoretical ambiguity [28, 29]. NPDGamma experiment measured the parity-violating asymmetry in polarized neutron capture on protons, $A_\gamma(\vec{n}\mathrm{p} \rightarrow \mathrm{d}\gamma)$, which is proportional to h_π^1. The experiment completed data collection and the final results show a small value for h_π^1 = [+2.6 ± 1.2 (stat.) ± 0.2 (sys.)] × 10^{-7} [29]. Other few body measurements include neutron capture on ^{3}He ($\vec{n}$+^{3}He) at the SNS and neutron spin rotation in ^{4}He ($\vec{n}$+^{4}He) at NIST (the focus of this proceeding). Within the framework of the DDH model, the $\vec{n}$+^{4}He is especially interesting. It is the isospin mirror of $\vec{p}$+^{4}He and it would allow an isoscalar/isovector separation. The linear combination of NN weak amplitudes in $\vec{n}$+^{4}He is almost orthogonal to the existing constraints from past measurements and the addition of $\vec{n}$+^{4}He gives strong constraints [30].

Recently, an alternative to the DDH potential has been developed based on pionless effective field theory (EFT) [31–33]. The advantages of this approach is that EFT treatment is model independent and has quantifiable errors and direct connection to QCD. When applying the $1/N_c$ expansion of QCD [34, 35], which is known to work at the ~10% level for deriving the relative strengths of the NN strong interacting couplings, the implications were very surprising. The first conclusion was that the isovector NN parity violation amplitudes are small compared to both isoscalar and isotensor NN weak interactions. The other implication was that the commonly used isoscalar/isovector plot to display constraints of NN weak interactions is very deceiving as it is hiding a dominant contribution from the isotensor amplitude [11].

In Ref. [11], Gardner, Holstein and Haxton have for the first time reorganized the theory analysis to take into account the recent theory results from the QCD $1/N_c$ expansion, which gives a natural scale for the relative size of the weak couplings. It leads to a completely different picture of the NN weak interaction. In particular, it has very important positive implications for the scientific impact of $\vec{n}$+^{4}He spin rotation. The theory is now in a position to predict $\vec{n}$+^{4}He neutron spin rotation from the Standard Model. The predicted size of the spin rotation angle is large, 6.8×10^{-7} rad/m, which is based on the assumption that the three N^{2}LO low-energy constants (LECs) are negligible [11]. However, very recent NPDGamma results [29] show that the N^{2}LO LECs can not be neglected. Including the N^{2}LO LECs and NPDGamma results [29] in the EFT+leading-N$_c$ expansion gives a neutron spin rotation angle in ^{4}He of $(4.4 \pm 1.5) \times 10^{-7}$ rad/m. The statistical error of the earlier version of the $\vec{n}$+^{4}He experiment was 9×10^{-7} rad/m [4], and with the new very intense beam of cold neutrons at NIST we should be able to reach a statistical accuracy of better than ~ 1×10^{-7} rad/m. The $\vec{n}$+^{3}He experiment is now completed at SNS and a final result with a very small value was released at the APS DNP2017 meeting. The $\vec{n}$+^{3}He collaboration is still investigating some of the associated systematic uncertainties. The $\vec{n}$+^{3}He experiment is sensitive to the isotensor component of the NN weak interaction, whereas $\vec{n}$+^{4}He is not. Since the isotensor NN weak amplitude is now understood from theoretical considerations to be one of the two leading-order terms in NN weak interaction, this is an important distinction between the two experiments.

In $\vec{n}$+^{4}He parity violation experiment, polarized slow neutrons are used to constrain hadronic weak coupling constants by measuring their parity violating spin rotation in liquid ^{4}He. The neutron spin rotation (NSR) collaboration measured a neutron spin rotation angle per unit length ($d\phi/dz$) of $[+1.7 \pm 9.1\,(\mathrm{stat.}) \pm 1.4\,(\mathrm{sys.})] \times 10^{-7}$ rad/m [4]. As mentioned previously, recent EFT+leading-N$_c$ calculations of $d\phi/dz$ showed a value of (4.4 ± 1.5) × 10^{-7} rad/m [11], therefore, an improvement on the statistical uncertainties to the level of 1 × 10^{-7} will be of a great value. At a minimum it would suffice to place a tight constraint on NN weak interaction amplitudes but can become the first test of the Standard Model in the NN weak sector if a nonzero value obtained as predicted by theory.

The $\vec{n}$+^{4}He parity-odd neutron spin rotation measurement involves passing a transversely polarized neutron through ^{4}He target and can be understood in terms of neutron optics. The parity-violating (PV) weak interaction causes the amplitudes of the positive and negative neutron helicity states moving through a medium to accumulate different phases. The difference ϕ_{PV} between the phase shifts of the helicity states leads to a rotation of the neutron polarization about its momentum, which manifestly violates parity [36]. The rotation angle per unit length is proportional to the forward limit of the parity-odd p-wave scattering amplitude f_{PV}, $d\phi/dz = 4\pi\rho f_{PV}/k$. Because f_{PV} is proportional to the parity-odd correlation $\vec{\sigma}_n \cdot \vec{k}_n$ with $\vec{\sigma}_n$ the neutron spin vector and $\vec{k}_n$ the neutron momentum, it tends to a constant for low energy neutrons in the absence of resonances [37]. The expected range of the parity-violating rotation angles for such few-body systems is of the order of 10^{-7} rad/m [9, 11, 38].

2 Experimental setup

Measuring a very small spin-rotation angle of 10^{-7} rad/m is very challenging and requires an apparatus that distinguishes the small parity-violating rotations from rotations

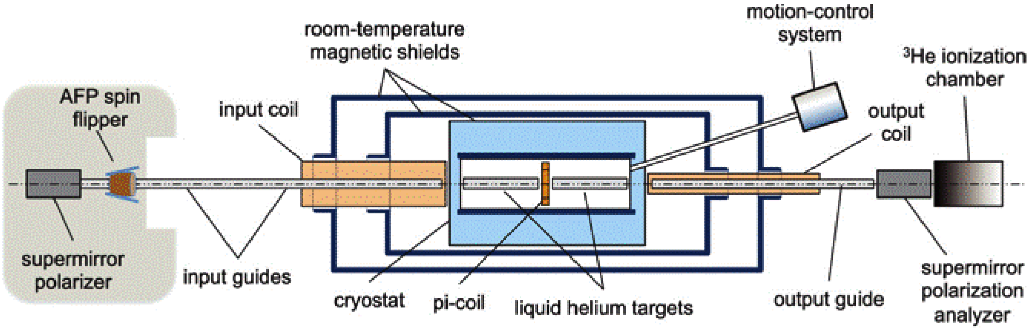


Figure 1. Top view diagram of the spin rotation apparatus [39]. The target consists of four chambers and a π-coil.

that arise from residual magnetic fields. The apparatus consists of an upgraded version of that used in the previous NSR measurement [40–42]. The old apparatus consisted of a neutron supermirror polarizer, an adiabatic RF spin-flipper, nonmagnetic neutron guides, spin transport, an adiabatic spin rotator, a neutron supermirror polarization analyzer, and a current-mode ion chamber [41, 42]. Figure 1 shows a diagram of the neutron beam-line and path along the spin rotation apparatus. Details of the measurement technique and the apparatus can be found elsewhere [39].

3 Toward an improved measurement

The recent high intensity neutron beam (NG-C) at NIST [43] has a factor of 20 increase on NG-6 flux which provides the opportunity to drastically improve the statistically limited previous NSR measurement [4]. However, NG-C has larger cross sectional area (11×11 cm^2) and larger divergence (5.95 mrad/Å) than the NG-6 neutron beam with 6×15 cm^2 cross sectional area and 2.0 mrad/Å divergence which necessitated upgrading the NSR apparatus to take advantage of the larger-area and more divergent NG-C beam.

In addition to matching the cross sectional area and divergence of NG-C beam, several other improvements on NSR setup were carried out. These improvements include two new 10×10 cm^2 polarizing super-mirrors with 60 Si/Gd $m = 2.5$ super-mirror (SM) blades to polarize and analyze the beam. The SM polarizer is shown in Figure 2. These SM benders have a transmission of greater than 90% for one spin state and a transmission of less than 0.5% for the other spin state.

To retain a greater faction of the more divergent beam on NG-C, new 10×10 cm^2 with $m = 2$ nonmagnetic (NiMo/Ti) super-mirror waveguides were constructed. The new input and output guides have $< 1\%$ depolarization probability per bounce. New input and output coils were built to match the larger guides. The coils consist of wires woven into grooves etched into hollow hexagonal plastic extrusions with a rectangular cavity in the center to fit the SM guides, as shown in Figure 3.

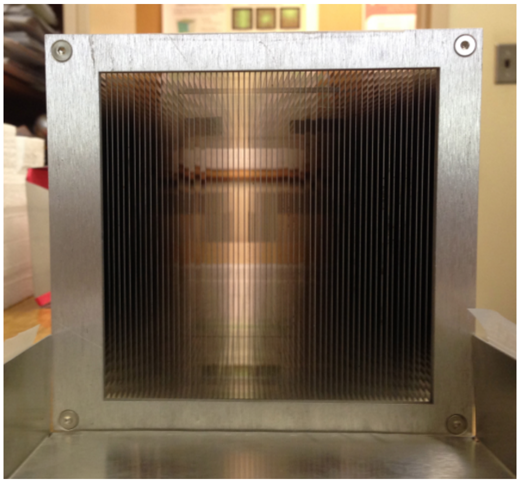


Figure 2. Front view of the SM. The image seen through the SM is deflected by 13 mrad.

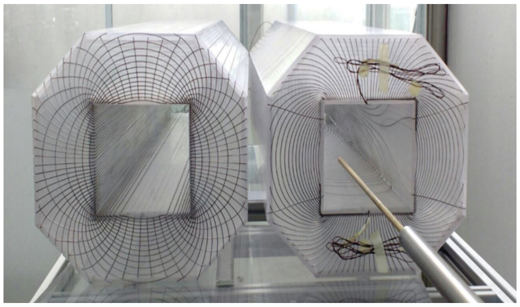


Figure 3. Front view of the input and output coils.

One of the main challenges during the previous NSR experiment was moving and maintaining liquid ^{4}He in the cryostat, and more than 25% of the beam-on time was lost in this process. Therefore, more efficient liquid ^{4}He pumping system that includes a ^{4}He liquefier was built. The ^{4}He

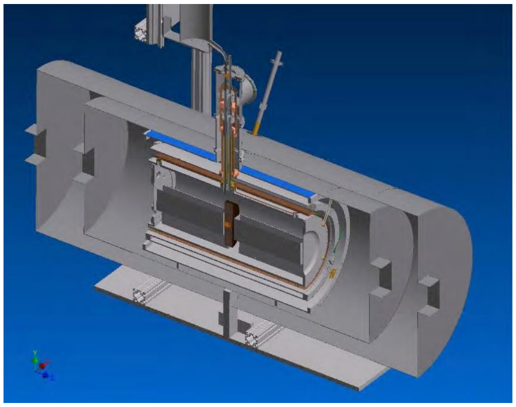


Figure 4. Cross sectional design view of the ^{4}He target.

liquefier is expected to operate continuously by converting helium boil-off gas from the helium cryostat to liquid helium and delivering liquid helium back into the cryostat. The ^{4}He liquefier was tested for three months of continuous operation at an observed liquefaction rate from warm gas of 12 liters per day. With automated operation capable of handling ~550 mW heat load, the ^{4}He liquefier will also reduce down time by improving the management of the liquid ^{4}He target. An enlarged ^{4}He target is being constructed to match the cross sectional area of the NG-C beam and fit with the new ^{4}He pumping system. Figure 4 shows the design of the ^{4}He target being constructed. All of the target components including the cryostat and ^{4}He liquefier are constructed and commissioned, and we are currently testing the ^{4}He pump.

A new 10 cm diameter ion chamber of similar design to the existing chamber was constructed. Additional improvements on the magnetic shielding in the target region were carried out to maintain magnetic fields below 10 μG in the target region which keeps the associated systematic uncertainties at $\sim 1.0 \times 10^{-7}$. Calculations of the involved systematic uncertainties are discussed elsewhere [39]. A summary of the improvements related to the new NSR experiment is shown in Table 1. With the higher NG-C flux and various upgraded experiment components, we expect the new experiment to reach statistical and systematic precision of [1.0 (stat) ± 1.0 (sys)] × 10^{-7} rad/m for the extracted parity violating spin rotation angle.

All of the NSR components except for the ^{4}He target were constructed and tested, and commissioned at LANSCE on FP12 beam-line. They were used for another measurement, that is discussed in section 4, and performed as they were intended.

4 Search for possible exotic spin dependent interactions using NSR apparatus

The NSR apparatus was also used to search for a possible new axial vector interaction in the millimeter to micron range using spin dependent interactions of neutrons with matter though exchange of spin 1 bosons as predicted in some extensions of the Standard Model. This experiment was performed on FP12 at the LANSCE facility at Los Alamos by passing transversely polarized slow neutrons through gaps between slabs of copper and float glass arranged so that the possible exotic interaction would tilt the plane of polarization along the neutron momentum [44]. The resulting rotation angle, ϕ = [+2.8 ± 4.6 (stat.) ± 4.0 (sys.)] × 10^{-5} rad/m, was consistent with zero [45]. This constraint improves on the previous upper bounds on g_A^2 by about three orders of magnitude for λ_c between 1 cm and 1 μm.

This result was obtained with a week worth of data at LANSCE and the recent high intensity neutron beam (NG-C) at NIST [43] provides the opportunity to drastically improve on these constraints. In addition, we are currently upgrading the room temperature target by using tungsten plates instead of copper plates to further increase mass density gradient in target which further enhances the studied effect. With one calendar month of beam time on NG-C and with the upgraded target we expect to set a new limit on possible exotic axial vector couplings of the neutron to matter which is more than two to three orders of magnitude more stringent than the previous measurement at LANSCE [45], as shown in Figure 5.

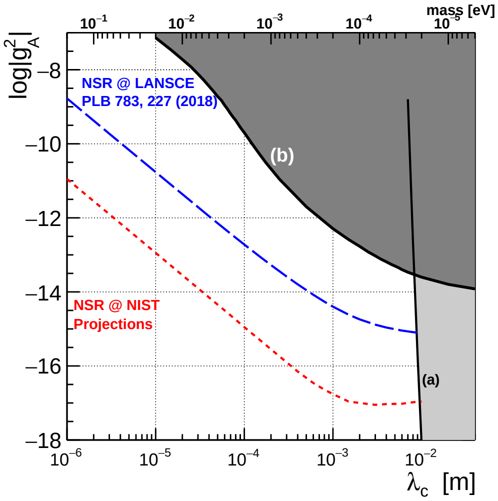


Figure 5. Projections of g_A^2 as a function of λ_c from the proposed measurement at NG-C (dashed-red) compared with constraints from LANSCE measurement (longdashed-blue) [45], K–^{3}He co-magnetometry (a) [46] and from a neutron measurement using Ramsey spectroscopy (b) [47].

5 Conclusion

Significant recent theoretical work predicted a relatively large neutron spin rotation, $(4.4 \pm 1.5) \times 10^{-7}$ rad/m, in ^{4}He without sensitivity to the isotensor component of the NN

Table 1. Summary of the upgraded and new components for NSR experiment on NG-C compared with those of NSR on NG-6.

Component	NSR on NG-6	NSR on NG-C
Counting Statistics	4.5×10^8	8×10^9
Polarizer/Analyzer		New / $m = 2.5$
Cross sectional area	5×5 cm^2	10×10 cm^2
Input/output guides	$m = 0.68$	$m = 2$
Duty factor		Reduced heat load
		Reduced fill/drain time
Background B-field	100 μG	10 μG
Uncertainties	9.1 (stat) ± 1.4 (sys)	1.0 (stat) ± 1.0 (sys)

weak interaction, a strong distinction between $\vec{n}$+^{3}He and $\vec{n}$+^{4}He. The NSR collaboration substantially improved the previous NSR apparatus to repeat the measurement at the high flux NG-C beam-line. This is expected to yield a measurement at the level of < [±1.0 (stat.) ± 1.0 (sys.)] × 10^{-7} rad/m which would provide the first test of the Standard Model in the NN weak sector.

The upgraded apparatus was used to make significant improvement in limits on spin-dependent fifth forces using a room temperature target. This constraint improves on the previous upper bounds on g_A^2 by 2–4 orders of magnitude for λ_c between 1 cm and 1 μm.

Acknowledgements

The work of M. Sarsour was supported by the US Department of Energy grant DE-SC0010443. E. Anderson, D. Esposito, I. Francis, K. Korsak, W. M. Snow, K. Steffen, and J. Vanderwerp acknowledge support from NSF PHY-1614545 and from the Indiana University Center for Spacetime and Symmetries. We acknowledge the support of the National Institute of Standards and Technology, US Department of Commerce, in providing some of the neutron facilities used in this work.

References

[1] J. Erler, M. Ramsey-Musolf, Prog. Part. Nucl. Phys. **54**, 351 (2005)

[2] M.J. Ramsey-Musolf, S.A. Page, Ann. Rev. Nucl. Part. Sci. **56**, 1 (2006)

[3] W.M. Snow, Eur. Phys. J. A **24**, 119 (2005)

[4] W.M. Snow et al., Phys. Rev. C **83**, 022501 (2011)

[5] C.S. Wood, S.C. Bennett, D. Cho, B.P. Masterson, J.L. Roberts, C.E. Tanner, C.E. Wieman, Science **275**, 1759 (1997)

[6] Y.B. Zeldovich, Sov. Phys. JETP **6**, 1184 (1957)

[7] V.V. Flambaum, I.B. Khriplovich, Sov. Phys. JETP **52**, 835 (1980)

[8] S. Tomsovic, M.B. Johnson, A.C. Hayes, J.D. Bowman, Phys. Rev. C **62**, 054607 (2000)

[9] B. Desplanques, J.F. Donoghue, B.R. Holstein, Annals Phys. **124**, 449 (1980)

[10] E.G. Adelberger, W.C. Haxon, Ann. Rev. Nucl. Part. Sci. **35**, 501 (1985)

[11] S. Gardner, W.C. Haxon, B.R. Holstein, Ann. Rev. Nucl. Part. Sci. **67**, 69 (2017)

[12] P. Eversheim et al., Phys. Lett. B **256**, 11 (1991)

[13] R. Balzer et al., Phys. Rev. Lett. **44**, 699 (1980)

[14] R. Balzer et al., Phys. Rev. C **30**, 1409 (1984)

[15] S. Kistryn et al., Phys. Rev. Lett. **58**, 1616 (1987)

[16] A.R. Berdoz et al., Phys. Rev. C **68**, 034004 (2003)

[17] A.R. Berdoz et al., **87**, 272301 (2001)

[18] J. Lang et al., Phys. Rev. Lett. **54**, 170 (1985)

[19] R. Henneck et al., Phys. Rev. Lett. **48**, 725 (1982)

[20] C.A. Barnes et al., Phys. Rev. Lett. **40**, 840 (1978)

[21] G. Ahrens et al., Nucl. Phys. A **390**, 486 (1982)

[22] M. Bini, T.F. Fazzini, G. Poggi, N. Taccetti, Phys. Rev. Lett. **55**, 795 (1985)

[23] S.A. Page et al., Phys. Rev. C **35**, 1119 (1987)

[24] E.G. Adelberger et al., Phys. Rev. C **27**, 2833 (1983)

[25] K. Elsener et al., Nucl. Phys. A **461**, 579 (1987)

[26] K. Elsener et al., Phys. Rev. Lett. **52**, 1476 (1984)

[27] W.S. Wilburn, J.D. Bowman, Phys. Rev. C **57**, 3425 (1998)

[28] M.M. Musgrave et al., Nucl. Instrum. Methods Phys. Res., Sect. A **895**, 19 (2018)

[29] D. Blyth et al. (2018), `1807.10192`

[30] Phys. Lett. B **125**, 1 (1983)

[31] S.L. Zhu, C. Maekawa, B. Holstein, M. Ramsey-Musolf, U. van Kolck, Nucl. Phys. A **748**, 435 (2005)

[32] L. Girlanda, Phys. Rev. C **77**, 067001 (2008)

[33] D.R. Phillips, M.R. Schindler, R.P. Springer, Nucl. Phys. A **822**, 1 (2009)

[34] D.R. Phillips, D. Samart, C. Schat, Phys. Rev. Lett. **114**, 062301 (2015)

[35] M.R. Schindler, R.P. Springer, J. Vanasse, Phys. Rev. C **93**, 025502 (2016)

[36] F.C. Michel, Phys. Rev. **133**, B329 (1964)

[37] L. Stodolsky, Nucl. Phys. B **197**, 213 (1982)

[38] B.R. Heckel, G.L. Greene, NBS Special Publication **711**, 90 (1986)

[39] W.M. Snow et al., Rev. Sci. Instrum. **86**, 055101 (2015)

[40] W. Snow, Nucl. Instrum. Methods Phys. Res., Sect. A **611**, 248 (2009)

[41] C. Bass et al., Nucl. Instrum. Methods Phys. Res., Sect. A **612**, 69 (2009)

[42] A. Micherdzinska et al., Nucl. Instrum. Methods Phys. Res., Sect. A **631**, 80 (2011)

[43] J.C. Cook, Rev. Sci. Instrum. **80**, 023101 (2009)

[44] C. Haddock et al., Nucl. Instrum. Methods Phys. Res., Sect. A **885**, 105 (2018)

[45] C. Haddock et al., Phys. Lett. B **783**, 227 (2018)

[46] G. Vasilakis, J.M. Brown, T.W. Kornack, M.V. Romalis, Phys. Rev. Lett. **103**, 261801 (2009)

[47] F.M. Piegsa, G. Pignol, Phys. Rev. Lett. **108**, 181801 (2012)

THERMAL DECOMPOSITION OF VEGETATIVE FUELS

Isaac T. Leventon [A] and Morgan C. Bruns [B]
[A] *National Institute of Standards and Technology, Fire Research Division,*
100 Bureau Drive; Building 224, Room A265; Gaithersburg, MD, 20899; United States

[B] *Virginia Military Institute, Department of Mechanical Engineering*
710 Nichols Hall; Lexington, VA 24450; United States

ABSTRACT

This manuscript presents new measurement data from milligram-scale thermal decomposition experiments - thermogravimetric analysis (TGA) and microscale combustion calorimetry (MCC) – conducted on stems and leaves of six plant species commonly found across the United States. For each fuel, measurement data from TGA experiments was analyzed to determine effective thermal decomposition mechanisms and the associated kinetics of their constituent reactions. MCC experiments were repeated under identical experimental conditions to determine the heats of complete combustion of all gaseous volatiles produced by these vegetative fuels and to validate the decomposition mechanisms and species char yields determined from TGA data. Through a coupled analysis of TGA and MCC measurement data**,** an estimate of the heats of combustion of the gaseous volatiles produced by individual reaction steps in the fuel's decomposition was also made. Between different fuels, distinct differences were measured in the onset temperature of decomposition, the temperature range of decomposition, the number of apparent reactions, and the peak measured mass loss and heat release rates (as well as the temperatures at which they occur). To analyze the impact of these variations on predictions of wildfire behavior, a modeling study was then conducted in which simulations of wildland fire experiments were repeated using the thermal decomposition mechanisms and heats of combustion determined for six of the fuel species tested in this work. Model-predicted fire spread rate in these simulations varied between 0.50 m s^{-1} and 1.09 m s^{-1}.

INTRODUCTION

With a growing number of people moving to areas in or near fire prone wildlands [1] and an increase in the number of large fires and total acres burned each year [2], wildland fires are an increasingly dangerous and costly problem. Accurate predictive modeling of current (or potential) uncontrolled wildland fires (i.e., quantitative prediction of fire intensity and spread rate) can mitigate the risk that these fires pose. Common models of wildland fire spread (e.g., BehavePlus [3] and FARSITE [4]) are relatively easy to use and can quickly provide fire spread predictions and deterministic assessments of fire hazard; however, they are based upon empirical relations (e.g., Rothermel [5]) defining a constant rate of spread for given conditions of slope, weather, wind, moisture, and a user-selected fuel model. These fuel models define representative physical parameters (surface area to volume ratio, particle size, and fuel bed depth), moisture content, and heat content (prescribed as 18.6 kJ g^{-1} for all but one, of 40 available, fuel models) [6]. Such empirically based models of flame spread have valuable applications (e.g., they are used for operational predictions of flame spread rate) but they do not incorporate the underlying processes controlling wildland fire spread behavior [7]. Thus, they are unable to predict transient fire behaviors and they may not be able to provide accurate predictions of fire behavior under changing ambient conditions or when a mixture of fuel sources (vegetative and structural) is present (e.g., at the wildland urban interface, WUI).

More powerful physics-based models (e.g., FDS [8]) can better capture the controlling mechanisms of wildland fires by solving governing equations for buoyant flow, heat and mass transfer, gas phase

Leventon, Isaac; Bruns, Morgan. "Thermal Decomposition of Vegetative Fuels." Presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, UK. July 01, 2019 - July 03, 2019.

combustion, and condensed phase thermal decomposition of fuels. These more capable simulation tools may be particularly valuable at the WUI, where simulation of the burning behavior of vegetative and structural fuels could be used to better inform structural and community wildfire resilience. Such models require a large number of inputs (e.g., fuel heat of combustion and radiative fraction, thermophysical properties of the vegetation and soil, and ambient conditions) to provide accurate predictions of wildland fire behavior. It has been shown that model predictions of the rate of spread of wildland fires are particularly sensitive to windspeed and the thermal decomposition temperature of the burning vegetative fuel [9]. Unfortunately, despite this sensitivity, comprehensive measurements of thermal decomposition are not readily available for a variety of common vegetative fuels and the fuel properties that are available from such experiments (i.e., the relevant data needed to parameterize physics-based models of wildfire spread) can be subject to large uncertainties [10].

Philpot conducted an early study on the pyrolysis of plant materials, using thermogravimetric analysis (TGA) and differential thermal analysis (DTA), to examine the relationship between a plant's mineral content and pyrolysis behavior (rate, onset temperature, and residue yields) [11]. Shafizadeh presented a thorough review of the pyrolysis of biomass, that described how the composition of its major components (cellulose, hemicellulose, and lignin) impacts its thermal properties and available decomposition pathways and that quantified the species yields of decomposition reactions how these vary with pyrolysis temperature [12]. Through the 1980s, Sussot led a series of studies on a broad range of wildland fuels, identifying the temperature range of decomposition for various plant components, the total heat needed for their pyrolysis, and the total energy released by combustion of these gaseous pyrolyzates (using an experimental apparatus that was a precursor to modern microscale combustion calorimetry) [13-16]. TGA experiments were also performed on multiple Mediterranean plant species to provide a ranking of their 'potential combustibility' [17] and other authors have thoroughly studied individual fuels (and their gaseous pyrolyzates) using multiple analytical methods [18]. Most recently, Amini and Safdari have characterized the char, tar, and gaseous species yields, and their respective chemical compositions, of the pyrolysis products of (live and dead samples of) fifteen species of vegetation native to the Southern United States [19, 20].

This manuscript presents new measurements from milligram-scale thermal decomposition experiments - thermogravimetric analysis (TGA) and microscale combustion calorimetry (MCC) – conducted on stems and leaves of six plant species commonly found in the United States. These fuels were collected in the summer of 2017 by the US Forest Service in Missoula, Montana (Lodgepole Pine, and Douglas-Fir) and in the North Mountain experimental area in Southern California (Chamise, Bigberry Manzanita, Desert Ceanothus, and Chaparral Whitethorn). All tests were conducted in nitrogen (i.e., in an anaerobic environment). Although the rate of degradation of these fuels may be affected by oxidation, it has been noted [21] that thin vegetative fuels will not ignite by pure radiative heating thus convective heating and flame 'bathing' is critical. Such direct flame impingement presents a 'fuel rich' (and thus a largely anaerobic) environment at the fuel's surface as it pyrolyzes.

For each fuel, sample mass and mass loss rate measured in TGA experiments were analyzed to determine effective thermal decomposition mechanisms and associated kinetics of these reactions. MCC experiments were repeated under the same experimental conditions to determine the heats of complete combustion of all gaseous pyrolyzates and species char yields. Additionally, by a coupled analysis of TGA and MCC measurement data**,** an estimate of the heats of combustion of the gaseous volatiles produced by individual reaction steps was made. To analyze the impact of variations in the degradation behavior of these fuels, a study was then conducted in which simulations of grassland fire experiments [8,9] were repeated using the unique thermal decomposition mechanisms and heats of combustion determined in this work. The sensitivity of model-predicted development of fire spread rate to these variations is discussed.

MATERIALS AND METHODS

Materials

The vegetative fuels studied in this work were obtained between May and July of 2017 from the United States Forest Service Pacific Southwest and Rocky Mountain Research Stations (which are located in Southern California and Western Montana, respectively). Species locations of origin, scientific names, and common names are provided in Table 1. The selected species represent vegetation types commonly found in the regions in which they were picked. For each species, a bulk sample – consisting of small branches with leaves attached – was picked from a series of randomly selected individual plants. Milligram-scale experiments conducted in this work were performed on both leaves and stems of all six plant species (except for Douglas-fir) thus creating a test matrix of eleven unique fuel species. Thermal analysis experiments were conducted on Douglas-fir leaves only.

Table 1. Vegetative fuels tested in this study

Origin	Scientific Name	Common Name
Pacific Southwest Research Station (North Mountain Experimental Area, California)	Adenostoma Fasciculatum	Chamise
	Arctostaphylos Glauca	Bigberry Manzanita
	Ceanothus Greggii	Desert Ceanothus
	Ceanothus Leucodermis	Chaparral Whitethorn
Rocky Mountain Research Station (Missoula, Montana)	Pinus Contorta	Lodgepole Pine
	Pseudotsuga Menziesii	Douglas-Fir

A preliminary series of TGA experiments was conducted on samples in one of two states: fresh (tested within 1-2 weeks after being picked) and after being microwaved three times (60 s each), sealed in plastic sample bags, and stored in a refrigerator. This treatment ensured the stability of samples (i.e., prevented decay, degradation, and/or molding) without affecting the chemical or physical structure of the foliage. Differences in measured sample mass and mass loss rate (MLR) of fresh and microwaved samples during these TGA experiments were negligible, thus all experimental measurements presented in this work were performed on samples that had been microwaved.

After this treatment, samples were cut into thin (< 0.75 mm thick) flat sections, less than 5 mm in length and between 4.5 mg – 6.5 mg in mass. For each test, leaves were kept whole/intact and stems were cut through their middle to create a flat surface that allowed for good thermal contact during experiments. 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, pressed flat into the base of alumina test crucibles, and weighed using a Mettler M3 analytical balance.

Thermal Analysis Experiments

Thermogravimetric Analysis (TGA) experiments were conducted 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 [22] - using a set of 6 pure metals, with melting points between 156.6 °C 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 vegetative fuel samples. All TGA experiments were conducted within three months of this calibration.

The temperature program used for TGA experiments included an initial heating at 10 °C min^{-1} to 75 °C followed by a 20-minute-long isotherm at that temperature, during which time the chamber was continuously purged with nitrogen. This conditioning period 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 °C min^{-1} to 700 °C (approximately 200 °C above the highest temperature at which a mass loss event was observed). Throughout this program, the test chamber was continuously purged with ultra-high purity (UHP) nitrogen at 50 mL min^{-1} to ensure thermal decomposition of samples occurred without oxidation. All tests were conducted in open alumina crucibles to allow gaseous pyrolyzates to escape unimpeded.

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 thermal analysis experiments. This baseline history (mass vs. temperature) was subtracted from the corresponding data obtained during experiments on vegetative fuel samples; all TGA measurement data presented in this work has been baseline-corrected in this manner. For each test, measured sample mass, m, was normalized by initial sample mass, m_0. Normalized sample mass loss rate $\frac{d(m/m_0)}{dt}$ [s^{-1}] was calculated as the numerical derivative of time-resolved sample mass curves and, prior to further analysis, noise in mass loss rate curves was reduced using a Savitzky-Golay filter. For each fuel species and sample type (stem and leaf), tests were repeated five times to accumulate necessary statistics; mass history curves from repeated experiments were averaged together prior to further analysis.

The relatively low heating rate (10 °C min^{-1}) used in these experiments was selected, in combination with the small sample masses used during testing, to ensure that samples did not experience significant temperature or composition gradients during heating [23,24]. Further, it has been demonstrated that an inverse analysis of total mass and mass loss rate data measured in TGA experiments conducted under these conditions can be used to determine effective reaction mechanisms, and associated reaction kinetics, that accurately describe the thermal decomposition of combustible solids [25].

Effective reaction mechanisms for all fuels were calculated assuming that measured decomposition behavior could be captured by a series of parallel, first order, Arrhenius rate reactions of the form

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -\sum_i (1-\nu_i) m_i A_i \exp\left(\frac{E_i}{RT}\right) \qquad [1]$$

where m is the total sample mass, m_i is the mass of component i, T is the sample temperature, R is the gas constant, and A_i and E_i are the kinetic parameters describing the reaction. Assuming a parallel reaction mechanism is reasonable in this case since it is likely that the vegetative fuels are composed of distinct components (cellulose, hemicellulose, and lignin) [12, 14]. TGA data does not allow for the determination of either the stoichiometric coefficient, ν_i, or the initial amount of component i present in the material, $m_{0,i}$. However, it is possible to determine the amount of mass lost as volatiles in each reaction from the TGA data. This quantity is simply $\Delta m_i \equiv m_{0,i}(1-\nu_i)$. The solid residue yield is related to the reaction mass losses through $\mu = 1 - \sum_i \Delta m_i$. For each of the fuels considered, the kinetic parameters A_i and E_i along with the reaction mass losses, Δm_i were determined using the algorithm developed in a recent work [26].

Microscale Combustion Calorimetry (MCC) experiments were conducted in an apparatus built in accordance with the relevant standard, ASTM D7309 [27]. In this test, specimens of known mass are thermally decomposed in an anaerobic environment at a constant heating rate. Gaseous volatiles released by pyrolyzing samples are mixed with an inert carrier gas and transported to a high temperature combustion chamber where they are forced to complete combustion in an oxygen rich environment. The heat released by combustion of these volatiles is computed from the rate of oxygen consumption in the gas stream exiting the combustion furnace. Sample temperature is continuously monitored during tests using a K-type thermocouple positioned directly beneath the sample crucible.

A temperature calibration was conducted as per best practices [27, 28] 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 vegetative fuel samples. All MCC tests were conducted within three months of this calibration. At the start of each day of testing, the MCC oxygen sensor was calibrated using a prepared gas mixture (19.19 % oxygen in nitrogen) and the total system calibration was checked using a reference material, polystyrene (Styron 665 GP). MCC experiments were conducted in accordance with ASTM D7309 [27] using 4.5 mg to 6.5 mg material samples that were pyrolyzed in UHP nitrogen. Samples were placed into open alumina crucibles, introduced into the pyrolysis chamber, and allowed to reach equilibrium at a temperature of 75 °C, at which point the chamber temperature was increased to 700 °C at a constant heating rate of 10 °C min^{-1}. Although the standard [27] recommends heating rates between 12 °C min^{-1} and 120 °C min^{-1}, a heating rate of 10 °C min^{-1} was selected in this study to provide measurements of sample heat release rate under comparable conditions to those used in TGA experiments.

The heat of complete combustion of all gaseous pyrolyzates ($\Delta H_{c,total}$) released by the pyrolyzing sample was determined as the integral of heat release rate, HRR, measured throughout the duration of tests divided by final volatilized mass, m_{vol} (i.e., initial sample mass minus the mass of char remaining after each test: $m_{vol} = m_0 - m_{char}$). Char yield, μ_{char}, was calculated by dividing m_{char} by m_0. Each fuel species was tested in the MCC at least three times to ensure reproducibility; $\Delta H_{c,total}$ and μ_{char} were calculated for each repeated experiment to accumulate necessary statistics. Values of $\Delta H_{c,total}$ and μ_{char} reported in this manuscript represent average values of repeated measurements.

Heats of complete combustion of the gases species produced during each reaction step, $\Delta H_{c,i}$, were determined by comparing heat release rate measured in MCC experiments and mass loss rate predicted by the decomposition model (which was developed on the basis on TGA experiments). The model was used to simulate sample mass loss rate under the conditions matching TGA and MCC experiments and a predicted heat release rate curve was generated by scaling the instantaneous rate of gaseous volatile production attributed to each reaction step by its corresponding $\Delta H_{c,i}$. These heats of combustion were adjusted in an iterative process until acceptable agreement between model-predicted and experimentally-measured heat release rate was obtained.

Numerical Simulations of Wildfire Spread

Computational Fluid Dynamics (CFD) simulations of wildfire spread were conducted in the NIST Fire Dynamics Simulator (FDS version 6.7.1) [29] to determine the sensitivity of flame spread rate predictions to measured variations in fuel decomposition behavior. Selected as a case study for this sensitivity analysis was a controlled burn of a 100 m by 100 m plot of kerosene grasslands conducted by the Commonwealth Scientific and Industrial Research Organization (CSIRO) of Australia between July and August of 1986 (Case C064) [30]. Measured properties of this case are reported in Table 2.

Table 2. Measured properties of CSIRO Grassland Fire Case C064 [30]

Property	Value
Wind Speed	4.6 m s^{-1}
Ambient Temperature	32 °C
Surface Area to Volume Ratio	9770 m^{-1}
Grass Height	0.21 m
Bulk Mass per Unit Area	0.283 kg m^{-2}
Moisture Fraction	6.3%

The computational domain in this case is 120 m by 120 m by 20 m. This domain is subdivided into 36 meshes, each with 0.5 m cubic grid cells. Increasing or decreasing grid size by a factor of 2 yields approximately a 5% deviation in model-predicted flame spread rate. In these simulations, Lagrangian particles are used to simulate blades of grass, which are modeled as slender cylinders whose diameters are inferred from the measured surface area to volume ratio. Each grid cell contains one simulated blade of grass; by applying a weighting factor, each explicitly modeled blade of grass represents approximately 5000 actual blades, thus matching the experimentally measured bulk mass per unit area. Blades of grass are rigidly fixed, perpendicular to the wind and the source of thermal radiation. Further detail on model assumptions concerning heat transfer and drag around blades of grass is provided elsewhere [9].

Wildfire simulations were repeated using the reaction mechanisms and associated kinetics and heats of combustion determined for six of the vegetative fuels tested in this work. Two additional simulations were also defined using decomposition models that represent the degradation of a typical leaf or stem; these models are referred to as 'Average Leaf' and 'Average Stem'. These cases provide insight into whether a given vegetative fuel's degradation mechanism can be estimated (based on existing knowledge) or if it must be uniquely measured to provide reasonable predictions of wildfire spread. Requisite parameters defining these decomposition models are reported, in further detail, in the 'Results and Discussion' section of this manuscript. All other relevant soil, vegetation, and combustion parameters used in these simulations are reported in Table 3; these values have been taken from a recent modeling study [9] and are typical of wood or cellulosic fuels. Ignition was defined to match experimental conditions, as described in a recent work [9].

Table 3. Assumed Fuel and Soil Properties for Wildfire Simulations [9]

Property	Value
Fuel Properties	
Chemical Composition	$C_6H_{10}O_5$
Radiative Fraction	0.35
Soot Yield	0.015
Specific Heat	1.5 kJ kg^{-1} K^{-1}
Conductivity	0.1 W m^{-1} K^{-1}
Density	512 kg m^{-3}
Heat of Pyrolysis	418 kJ kg^{-1}
Soil Properties	
Soil Specific Heat	2.0 kg^{-1} K^{-1}
Soil Conductivity	0.25 W m^{-1} K^{-1}
Soil Density	1300 kg m^{-3}

RESULTS AND DISCUSSION

Thermal Analysis Experiments

Figure 1 plots the results of TGA experiments conducted on stem and leaf samples of all plant species tested in this work. Solid black lines represent the mean of repeated experiments; the shaded area is calculated as one standard deviation. Also shown in Fig. 1, as red lines, are model-predictions of sample behavior during TGA experiments (dashed and dotted lines represent total and reaction-step-specific model-predicted residual mass loss rates, respectively). Optimized kinetic parameters for each of these reaction mechanisms are provided in Table 4. The details of this fitting process are provided in a related work [26]. As seen in Fig. 1, all models capture experimentally measured mass loss rate data with reasonable accuracy.

Table 4. Kinetic Parameters Describing Decomposition of Vegetative Fuels

Sample Name	A_1 (s^{-1})	E_1 (kJ $kmol^{-1}$)	Δm_1	A_2 (s^{-1})	E_2 (kJ $kmol^{-1}$)	Δm_2	A_3 (s^{-1})	E_3 (kJ $kmol^{-1}$)	Δm_3
Leaves									
Chamise	9.98×10^{2}	5.67×10^{4}	0.30	1.21×10^{4}	7.69×10^{4}	0.32	3.39×10^{8}	1.37×10^{5}	0.11
Bigberry Manzanita	2.12×10^{3}	5.91×10^{4}	0.34	3.64×10^{9}	1.39×10^{5}	0.12	1.07×10^{3}	7.23×10^{4}	0.27
Desert Ceanothus	2.22	3.32×10^{4}	0.64	9.99×10^{10}	1.52×10^{5}	0.01	3.80×10^{14}	2.14×10^{5}	0.02
Chaparral Whitethorn	9.85×10^{3}	6.68×10^{4}	0.17	1.15×10^{5}	8.68×10^{4}	0.24	1.53	4.36×10^{4}	0.21
Lodgepole Pine	2.38×10^{5}	8.99×10^{4}	0.38	2.85×10^{8}	1.12×10^{5}	0.11	6.39×10^{1}	5.67×10^{4}	0.23
Douglas-Fir	3.35×10^{4}	7.07×10^{4}	0.24	1.45×10^{7}	1.09×10^{5}	0.26	2.13×10^{1}	5.17×10^{4}	0.26
Stems									
Chamise	9.56×10^{6}	1.08×10^{5}	0.38	3.40×10^{12}	1.58×10^{5}	0.07	1.43×10^{2}	6.04×10^{4}	0.24
Bigberry Manzanita	4.85×10^{5}	7.41×10^{4}	0.10	2.14×10^{6}	1.01×10^{5}	0.53	2.06×10^{14}	1.79×10^{5}	0.07
Desert Ceanothus	1.16×10^{8}	1.20×10^{5}	0.64	5.05×10^{10}	1.39×10^{5}	0.13			
Chaparral Whitethorn	3.23×10^{9}	1.26×10^{5}	0.07	9.56×10^{5}	9.86×10^{4}	0.69			
Lodgepole Pine	3.45×10^{6}	8.91×10^{4}	0.30	5.97×10^{7}	1.15×10^{5}	0.22	1.51×10^{1}	4.96×10^{4}	0.26
Average Stem*	8.58×10^{5}	9.64×10^{4}	0.49	1.03×10^{16}	1.95×10^{5}	0.07			
Average Leaf*	1.22×10^{3}	5.75×10^{4}	0.23	2.46×10^{5}	9.03×10^{4}	0.23	1.32×10^{2}	6.02×10^{4}	0.25

* Effective values representing the thermal decomposition of a typical leaf or stem tested in this work

Leventon, Isaac; Bruns, Morgan. "Thermal Decomposition of Vegetative Fuels." Presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, UK. July 01, 2019 - July 03, 2019.

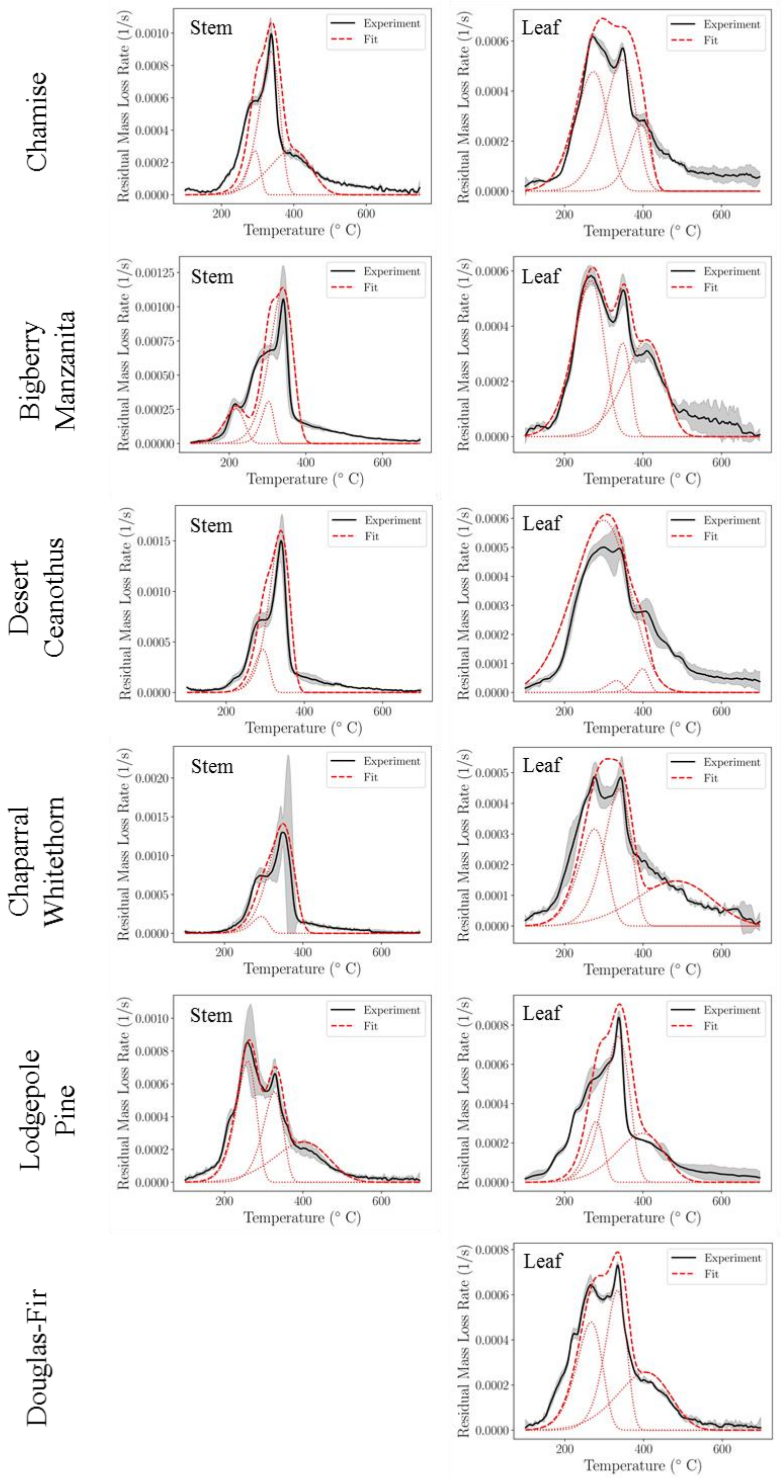


Figure 1. Experimentally measured and model predicted mass loss rate data of stem (left) and leaf (right) samples of Chamise, Bigberry Manzanita, Desert Ceanothus, Chaparral Whitethorn, Lodgepole Pine, and Douglas-Fir in TGA tests conducted in Nitrogen at 10 °C min^{-1}.

Temperature-resolved mass loss rate measurements obtained during TGA experiments on all vegetative fuels tested in this work are plotted together in Fig. 2, to allow for a qualitative comparison of the thermal degradation behavior of each of these fuels. As seen here, between each fuel, there are distinct differences in the onset temperature of degradation, the number of apparent reactions, and the peak measured mass loss rate (as well as the temperature at which it occurs). Across all samples tested, the temperature corresponding to each of these reaction peaks varies between 220 and 485 °C. In general, two reaction peaks (local maxima in $\frac{d(m/m_0)}{dt}$) are observed for stem samples. Stem samples demonstrated higher peak mass loss rates than leaves and their decomposition generally occurred over a narrower temperature range, with little mass loss above 400 °C. Leaf samples were characterized by a series of overlapping reactions, typically at least three, that occurred over a wider temperature range. These results support previous observations by Sussot [14], who noted that the relative proportions of extractives (i.e., volatile hydrocarbons), hemicellulose, cellulose, and lignin appear to explain the mg-scale thermal degradation behavior of typical forest fuels. Also plotted in this figure (black lines) are curves representing the degradation behavior of an idealized ‘Average’ stem or leaf. It should be stressed that these curves do not represent actual measurement data: rather, they define a curve representative of typical stem or leaf degradation, which is used in this work to help interpret wildfire simulation results. Kinetic parameters used to define these curves are reported in Table 4 under the sample names ‘Average Stem’ and ‘Average Leaf’.

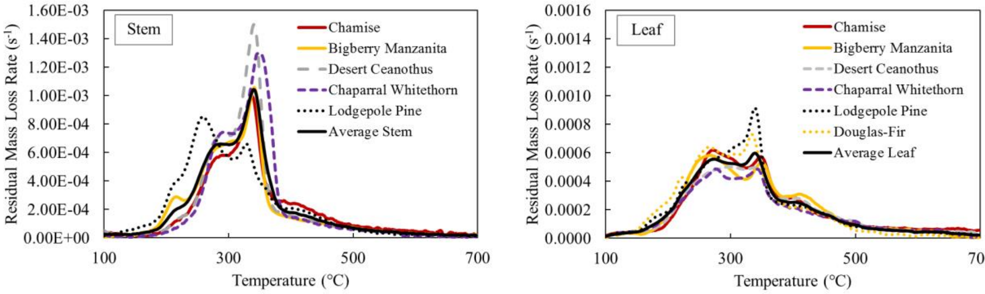


Figure 2. Measured mass loss rates of all vegetative fuels tested in this work when heated (in ultra-high purity nitrogen) at 10 °C min^{-1} in the TGA. Each curve represents an average of five repeated experiments; for clarity, measurement uncertainty is not plotted (this is shown in Fig. 1).

Figure 3 plots experimentally measured heat release rate, HRR, (normalized by volatilized sample mass, m_{vol}) of stem and leaf samples of all plant species tested in this work in the MCC. Experimental measurements are plotted as solid black lines. Measured values of $\Delta H_{c,total}$ and μ_{char} are reported in Table 5. Uncertainties are reported for μ_{char} as one standard deviation and for $\Delta H_{c,total}$ as the maximum of either: one standard deviation of repeated measurements or 5% of the average value (which represents the inherent uncertainty in oxygen consumption measurements of heat release [31]). Measured values of $\Delta H_{c,total}$ vary between 8.9 14.4 kJ g^{-1} and 14.4 kJ g^{-1}, which is a significant deviation from the value of 18.6 kJ g^{-1} that is prescribed for most (39 of 40) fuels in the Standard Fire Behavior Fuel Models [6]. For the fuels tests in this work, exclusive of Lodgepole Pine stems, measured values of $\Delta H_{c,total}$ are, on average, 17% greater for leaves than for stems. The heat of combustion of Lodgepole Pine stems is neglected in this comparison because it is, 42% greater than the average $\Delta H_{c,total}$ of the other stems tested in this work.

Also shown in Fig. 3 are model predictions of heat release rate (i.e., model-predicted mass loss rate, scaled by the heats of combustion, $\Delta H_{c,i}$). Optimized values for $\Delta H_{c,i}$ are presented in Table 5; these values are calculated based on the ratio of experimentally measured heat release rate (in MCC tests) and mass loss rate (in TGA tests) at the peak mass loss rate of each of the reaction steps defined by the fuel’s decomposition model (Table 4). This calculation accounts for relative fraction of gaseous volatiles produced by each reaction step at the temperature of interest. The uncertainty in reported values of $\Delta H_{c,i}$ is estimated to be 15% based on a propagation of the 5% uncertainty in heat release measurements and the differences between of model-predicted and experimentally-measured sample mass loss rate (see Fig. 1), which was used to generate the HRR curves shown in Fig. 3.

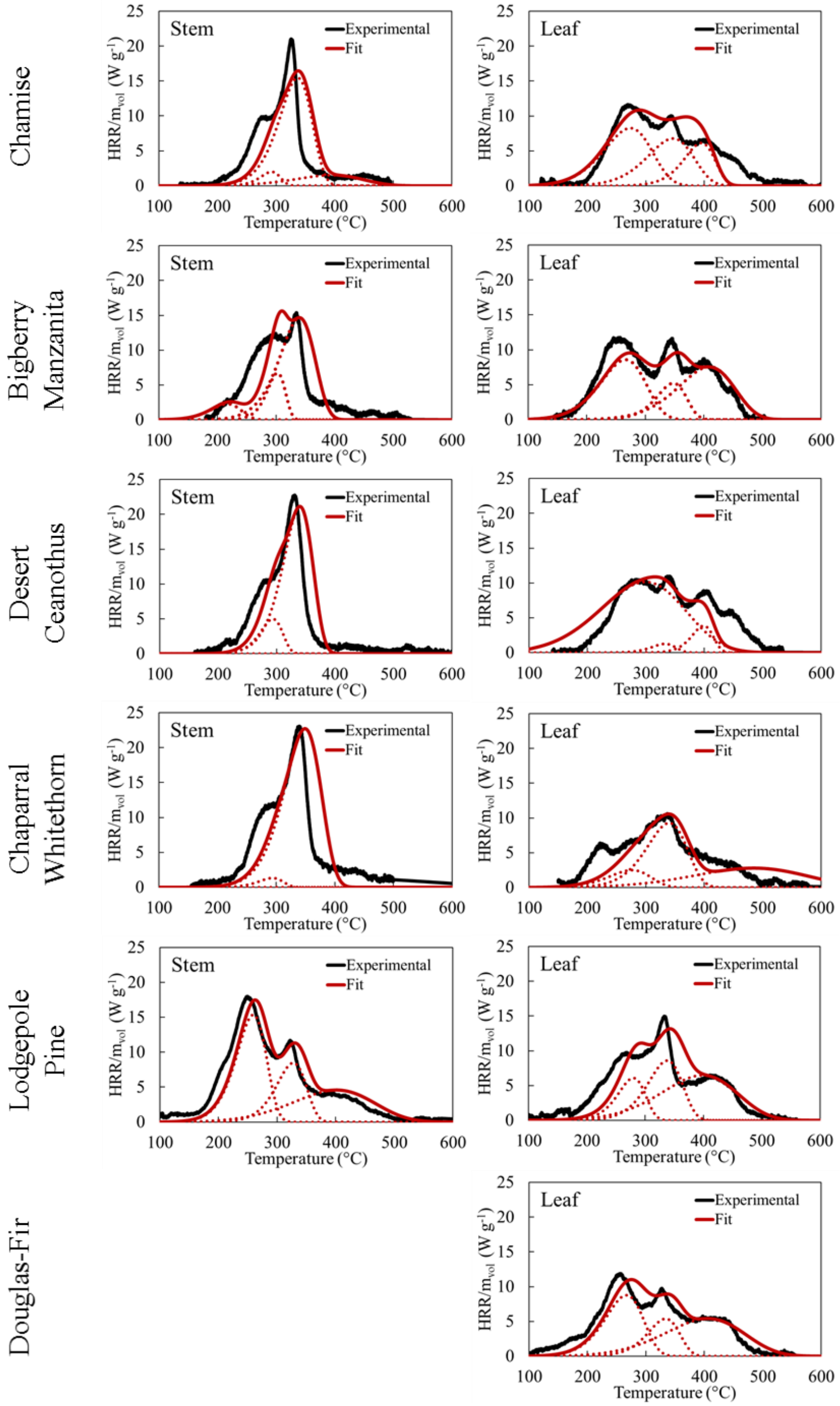


Figure 3. Experimentally-measured and model-predicted heat release rate of stem (left) and leaf (right) samples of Chamise, Bigberry Manzanita, Desert Ceanothus, Chaparral Whitethorn, Lodgepole Pine, and Douglas-Fir in MCC tests conducted in Nitrogen at 10 °C min^{-1}.

Table 5. Heats of Complete Combustion and Char Yields of Vegetative Fuels

Sample Name	$\Delta H_{c,1}$ (kJ g^{-1})	$\Delta H_{c,2}$ (kJ g^{-1})	$\Delta H_{c,3}$ (kJ g^{-1})	$\Delta H_{c,total}$ (kJ g^{-1})	μ_{char} (-)
Leaves					
Chamise	17.3±2.6	12.9±1.9	22.9±3.4	11.7±1.2	0.25±0.04
Bigberry Manzanita	15.4±2.3	14.9±2.2	21.8±3.3	12.4±0.9	0.22±0.06
Desert Ceanothus	17.1±2.6	30.7±4.6	47.4±7.1	12.3±1.1	0.32±0.03
Chaparral Whitethorn	7.9±1.2	20.7±3.1	19.0±2.8	10.4±1.8	0.33±0.04
Lodgepole Pine	10.8±1.6	16.1±2.4	23.6±3.5	12.6±0.6	0.24±0.04
Douglas-Fir	18.3±2.7	8.7±1.3	21.1±3.2	12.2±0.6	0.25±0.04
Average Leaf *	-	-	-	11.9±0.8	0.27±0.05
Stems					
Chamise	17.5±2.6	7.3±1.1	5.5±0.8	8.9±0.6	0.27±0.04
Bigberry Manzanita	9.0±1.4	12.9±1.9	21.9±3.3	8.9±0.9	0.37±0.06
Desert Ceanothus	13.2±2.0	11.7±1.8		9.1±0.5	0.25±0.06
Chaparral Whitethorn	6.1±0.9	16.1±2.4		11.5±2.6	0.23±0.05
Lodgepole Pine	20.8±3.1	15.4±2.3	18.6±2.8	14.4±2.0	0.22±0.04
Average Stem*	-	-	-	10.9±2.3	0.27±0.05

*Calculated as the mean value of $\Delta H_{c,total}$ or μ_{char} measured for all stem or leaf species

Numerical Simulations of Wildfire Spread

Figure. 4 plots time-resolved predictions of fire front location in FDS simulations using the decomposition models developed in this work for stems and leaves of Bigberry Manzanita, Chamise, and Chaparral Whitethorn. The fire front is defined as the location of the maximum gas temperature in a 1 m wide, 1 m tall strip along the centerline of the grass field. As seen in Fig. 4, propagation of this fire front across the length of the field occurred at a fairly constant rate, R, which varied between 0.50 and 1.09 m s^{-1} (factor of two difference) when using the decomposition models developed in this work.

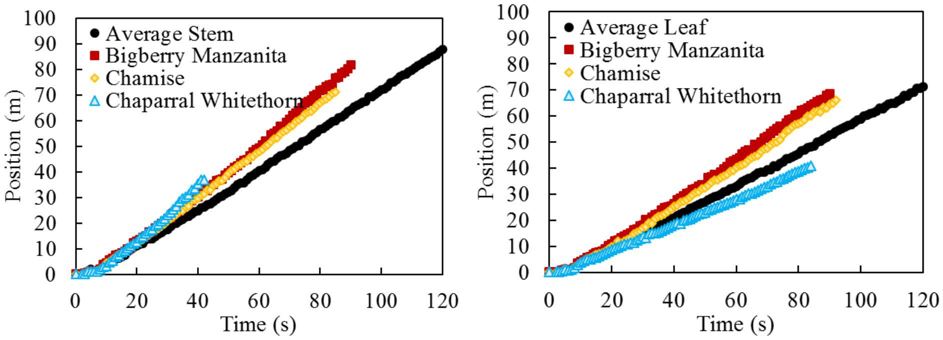


Figure 4. Comparison of predicted fire front position of CSIRO C064 Grassland Fires simulated in FDS 6.7.1 using the thermal decomposition models developed for three stem (left) and leaf (right) vegetative fuels tested in this work.

The calculated fire spread rates for 'Average Stem' and 'Average Leaf' models are R_{avg}^{stem} = 0.75 m s^{-1} and R_{avg}^{leaf}= 0.61 m s^{-1}, respectively (a relative difference of 24%). This indicates that measured differences in the decomposition behavior of stems and leaves (e.g., in heats of combustion or decomposition temperature, range, and peak mass loss rate) produce distinct differences in the global behavior of wildfires in FDS simulations. To better understand how well average stem or leaf models capture the behavior of fuel-specific stem or leaf models, a simple sensitivity analysis was performed. Model sensitivity of predicted fire spread rate was calculated as $S = \frac{(R_i - R_{avg})}{R_{avg}}$ where R_i and R_{avg} represent the fire spread rates calculated using a specific (stem or leaf) fuel model or that of the average (stem or leaf) fuel model. Calculated sensitivity varied between 0.22 and 0.45 for the three stem species and -0.17 and 0.37 for the three leaf species simulated here. This indicates that, not only is the predicted wildfire spread rate of leaves different from that of stems, but there are distinct differences between the predicted fire spread rates of individual fuels of the same type.

CONCLUSIONS

In this work, the thermal degradation behavior of stem and leaf samples of six vegetative fuels commonly found in the United States was examined through a series of thermogravimetric analysis (TGA) and microscale combustion calorimetry (MCC) experiments. Measurement data from TGA experiments was used to determine effective thermal decomposition mechanisms – consisting of a series of parallel, first order, Arrhenius rate reactions – and associated kinetics. MCC experiments were repeated under the same experimental conditions as TGA tests, thus allowing for the determination the heats of complete combustion of all gaseous pyrolyzates released by the degrading sample, $\Delta H_{c,total}$, and heats of complete combustion of the gases species produced during each reaction step, $\Delta H_{c,i}$. MCC data was also used for the validation of the decomposition mechanisms and species char yields, μ_{char}, determined from TGA experiments.

Distinct differences were measured in the onset temperature of degradation, the number of apparent reactions, and the peak measured mass loss and heat release rates (as well as the temperatures at which they occur). Across all samples tested, the temperatures corresponding to the peaks of each reaction step in the determined degradation mechanisms varied by more than 250 °C. In general, stem samples demonstrated higher peak mass loss rates than leaf samples and their decomposition generally occurred over a narrower temperature range, with little mass loss above 400 °C. Leaf samples were characterized by a series of overlapping reactions that occurred over a wider temperature range and generally had higher heats of combustion. FDS simulations were run to examine the sensitivity of model predictions of wildfire spread rate to these measured variations in the thermal decomposition behavior of these fuels. Six fuel models were selected for this sensitivity analysis: simulations demonstrated a clear dependence on fuel decomposition mechanism, with predictions of wildfire spread rate varying between 0.50 m s^{-1} and 1.09 m s^{-1}.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Sara McCallister of the Rocky Mountain Research Station and Dr. David Weise of the Pacific Southwest Research Station for collecting and sharing the vegetative fuel samples tested in this work and for helpful discussions during the planning stages of this project. Preliminary TGA experiments and related analysis were conducted at NIST by David Hoddinott; the authors are grateful for his assistance. Wildfire simulations and a final review of this manuscript were assisted by Dr. Kevin McGrattan of NIST – this was greatly appreciated, especially given the condensed timeframe in which this help was offered.

REFERENCES

1. Government Accountability Office (GAO), “Technology Assessment: Protecting Structures and Improving Communications During Wild-Land Fires,” Technical Report GAO-05-380, United States Government Accountability Office: Washington, DC. (2005)
2. National Interagency Fire Center (NIFC). Total wildland fires and acres, 1983–2017. Accessed October 2018. www.nifc.gov/fireInfo/fireInfo_stats_totalFires.html.
3. Heinsch, F. A., Andrews, P.L., “BehavePlus fire modeling system version 5.0: Design and Features,” Gen. Tec. Rep., RMRS-GTR-249, US Department of Agriculture, Rocky Mountain Research Station, Fort Collins, CO. (2010)
4. Finney, M. A., “FARSITE: Fire Area Simulator – Model Development and Evaluation,” RMRS-RP-4, US Department of Agriculture, Rocky Mountain Research Station, For Collins CO. (2004)
5. Rothermel, R. C., “A Mathematical Model for Predicting Fire Spread in Wildand Fuels,” USDA Forest Service, Intermountain Forest and Range Experiment Station, Research Paper INT-115. (1972)

6. Scott, J.H., Burgan, R.E., "Standard Fire Behavior Fuel Models: A Comprehensive Set for Use with Rothermel's Surface Fire Spread Model," RMRS-GTR-153, US Department of Agriculture, Rocky Mountain Research Station, Fort Collins, CO. (2005)
7. Finney, M. A., Cohen, J.D., McAllister, S.S., Jolly, M., "On the Need for a Theory of Wildland Fire Spread," International Journal of Wildland Fire 22: p. 25-36. (2013)
8. Mell, W., Jenkins, M.A., Gould, J., Cheney, P., "A physics-based approach to modeling grassland fires," International Journal of Wildland Fire 16: p. 1-22. (2007)
9. McGrattan, K.B., "Progress in Modeling Wildland Fires using Computational Flud Dynamics," 10th US Combustion Meeting, College Park, MD. (2017)
10. Gollner, M., Trouvé, A., "Towards Data-Driven Operational Wildfire Spread Modeling," WIFIRE Workshop, San Diego, CA. (2015)
11. Philpot, C.W., "Influence of Mineral Content on the Pyrolysis of Plant Materials," Forest Science 16: p. 461-471. (1970)
12. Shafizadeh, F., "Introduction to Pyrolysis of Biomass," Journal of Analytical and Applied Pyrolysis 3: p. 283-305. (1982)
13. Sussot, R. A., "Thermal Behavior of Conifer Needle Extractives," Forest Science 26: p. 347-360. (1980)
14. Sussot, R. A., "Characterization of the Thermal Properties of Forest Fuels by Combustible Gas Analysis," Forest Science 28: p. 404-420. (1982)
15. Sussot, R. A., "Differential Scanning Calorimetry of Forest Fuels," Forest Science 28: p. 839-851. (1982)
16. Rogers, J.M., Sussott, R.A., Kelsey R.G., "Chemical Composition of Forest Fuels Affecting Their Thermal Behavior," Canadian Journal of Forest Research 16: p. 721-726. (1986)
17. Dimitrakopoulos, A.P., "Thermogravimetric Analysis of Mediterranean Plant Species," Journal of Analytical and Applied Pyrolysis 60: p. 123-130. (2001)
18. Statheropoulos, M., Liodakis, S., Tzamtzis, N., Pappa, A., Kyriakou, S., "Thermal Degradation of Pinus Halepensis Pine-Needles Using Various Analytical Methods," Journal of Analytical and Applied Pyrolysis 43: p. 115-123. (1997)
19. Safdari, M-S., Rahmati, M., Amini, E., Howarth, J.E., Berryhill, J.P., Dietenberger, M., Weise, D.R., Fletcher, T.H., "Characterization of pyrolysis products from fast pyrolysis of live and dead vegetation native to the Southern United States," Fuel 229: p. 151-166. (2018)
20. Amini, E., Safdari, M-S., DeYoung, J.T., Weise, D.R., Fletcher, T.H., "Characterization of pyrolysis products from slow pyrolysis of live and dead vegetation native to the Southern United States," Fuel 235: p. 1475-1491. (2019)
21. McAllister, S., Finney, M., "Convection Ignition of Live Forest Fuels," Fire Safety Science 11: p. 1312-1325. (2014)
22. NETZSCH, "Software Manual (STA 449 F1 & F3) Temperature and Sensitivity Calibration," Wittelsbacherstrasse 42, 95100 Selb, Germany: NETZSCH Gerätebau GmbH. (2012)
23. Lyon, R.E., Safronava, N., Senese, J., Stoliarov S.I., "Thermokinetic model of sample response in nonisothermal analysis," Thermochimica Acta 545: 82-89. (2012)
24. Vyazovkin, S., Chrissafis, K., Di Lorenzo, M.L., Koga, N., Pijolat, M., Roduit, B., Sbirrazzuoli, N. Sunol, J.J., "ICTAC Kinetics Committee Recommendations for Performing Kinetic Computations on Thermal Analysis Data," Thermochimica Acta 590: p. 1-23. (2014)
25. Stoliarov S.I., Li J., "Parameterization and Validation of Pyrolysis Models for Polymeric Materials; Fire Technology 52: p. 79-91. (2016)
26. Bruns, M.C., Leventon, I.T., "Automated Fitting of Thermogravimetric Analysis Data," Interflam 2019
27. ASTM D7309, "Standard Test Method for Determining Flammability Characteristics of Plastics and Other Solid Materials Using Microscale Combustion," ASTM International: West Conshohocken, PA, USA. (2013)
28. Lyon, R. E., Walters, R. N., Stoliarov, S. I., Safronava, N., "Principles and Practice of Microscale Combustion Calorimetry," FAA Report, DOT/FAA/TC-12/53 R3 (April 2013)
29. McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., Vanella, M., "Fire Dynamics Simulator Technical Reference Guide," NIST Special Publication 1018-1, Sixth Edition. (2019)
30. Cheney, N.P., Gould, J.S., Catchpole, W.R., "The Influence of Fuel, Weather and Fire Shape Variables on Fire-Spread in Grasslands," International Journal of Wildland Fire 3: p. 31–44, 1993.
31. Hugget, C. "Estimation of Rate of Heat Release by Means of Oxygen Consumption Measurements," Fire and Materials 4: p. 61-65. (1980)

STANDARD LEDS WITH SUPERIOR LONG-TERM STABILITY

Yuqin Zong[1], Weiqiang Zhao[2], C. Cameron Miller[1]
[1] National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA
[2] Guest researcher at NIST from National Institute of Metrology (NIM), Beijing, CHINA
yuqin.zong@nist.gov

Abstract

A large-chip standard LED with superior long-term stability have been developed at NIST. The standard LED uses a large, specialty die rated for 50 W but is operated under 3 W to eliminate aging effect. The standard LED was seasoned for one year and measured for its long-term stability for three years. The measurement result shows that the long-term stability is on the level of 0.1 % in the three-year time. The LED can be used as a transfer standard for luminous intensity, luminance, and total luminous flux.

Keywords: Standard LED, Long-term Stability, Photometry.

1 Introduction

Conventional incandescent standard lamps for detector and instrument calibrations (e.g., CIE Standard Illuminant A lamps used for photometer calibrations) are mostly discontinued by the manufacturers due to the revolution of the lighting technology in recent years. To address this issue and also to take advantages of the new solid-state lighting technology, the International Commission on Illumination (CIE) established a new technical committee TC 2-90 recently for developing new reference spectrum based on white LEDs for photometer calibrations. In addition, standard LEDs are being developed by LED manufacturers, national metrology labs, and instrument manufacturers. Most of these standard LEDs use a typical LED package design. They are temperature-controlled and have good short stabilities (days). However, on-shelf long-term stability of the existing standard LEDs is not well known and is typically not specified by the manufacturers, which limits their uses only as short-term working standards.

2 The new standard LEDs

We developed a standard LED using a large, specialty die for improving long-term stability. A schematic of the LED package is shown in Figure 1. The large die is a 3 mm × 3 mm square monolithic device and is rated for 13.5 A, 50 W continuous operation. The die is packaged into a protection housing with a glass window to keep it clean and dry (the same form factor as that of a photodiode). The LED package is mounted on a star shape metal core printed circuit board (MCPCB). The light-emitting surface of the die is flat and coated with a thin layer of phosphor (with no encapsulation). The LED package has a uniform luminance distribution and a near Lambertion beam pattern. These characteristics make it a good luminous intensity transfer standard and a luminance transfer standard in addition to a good total luminous flux transfer standard. This type of LEDs is often used for applications such as high-power back light of an LCD projector, high-power fiber-coupled illumination, etc. The luminous efficacy of this type of specialty LEDs is low compared to a typical white LED, which is not an issue to be used as a standard LED.

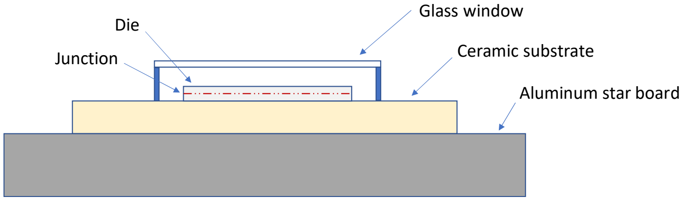


Figure 1 – Illustration of the LED package

Zong, Yuqin; Zhao, Weiqiang; Miller, C Cameron. "Standard LEDS with Superior Long-Term Stability." Presented at 29th Quadrennial Session of the International Commission on Illumination (CIE), Washington, DC, US. June 14, 2019 - June 22, 2019.

The LED package is mounted on the cold plate of a compact (50.8 mm × 50.8 mm × 58 mm) temperature-controlled mount (TCM). A 0.2 mm thick graphite-based thermal interfacing sheet is used between the star board and the cold plate to minimize the thermal resistance between the junction and the cold plate. The TCM has two connectors for operating the LED package; one for a current source and the other for a temperature controller. The housing of the TCM is clear anodized for reducing self-absorption when it is used in an integrating sphere. The LED's thermal resistance from the junction to the TCM cold plate is low (approximately 1 °C/W as opposed to more than 10 °C/W for a typical LED package) and the TCM is set to operate at the room temperature (25 °C), which minimizes the influence from ambient air temperature and air movement. Further, the 50 W LED package is operated only at 3 W level, which virtually eliminates its aging effect. A photograph of a standard LED including a LED package and a TCM is shown in Figure 2.

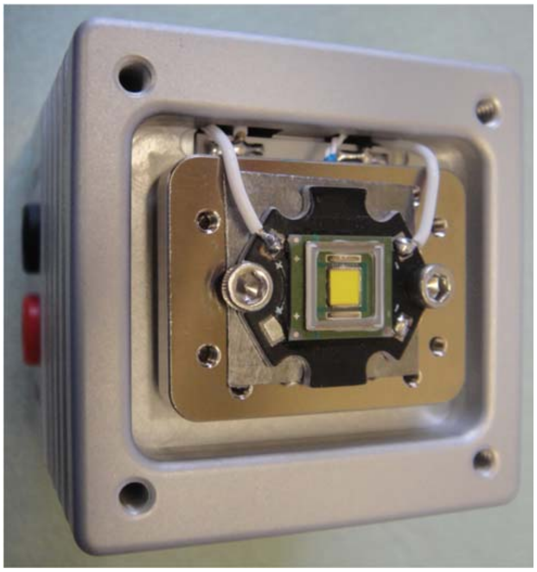


Figure 2 – Photograph of a standard LED

The LED is operated using a compact current source and temperature controller combo unit. It has a useful LED protection feature so that if its temperature controller is turned off accidently the current source will be automatically turned off immediately.

3 Stability of the standard LED

The standard LED was first evaluated for repeatability. A highly stable, temperature-controlled photometer was set up for measuring an standard LED at a distance of 400 mm for luminous intensity. The standard LED was operated at 1 A with its TCM temperature controlled at 25 °C. It was measured 10 times for luminous intensity. Between two consecutive measurements, the LED was turned off for at least one hour. Figure 3 shows the repeatability of the LED, which is better than 0.01 % after it is stabilized for one hour.

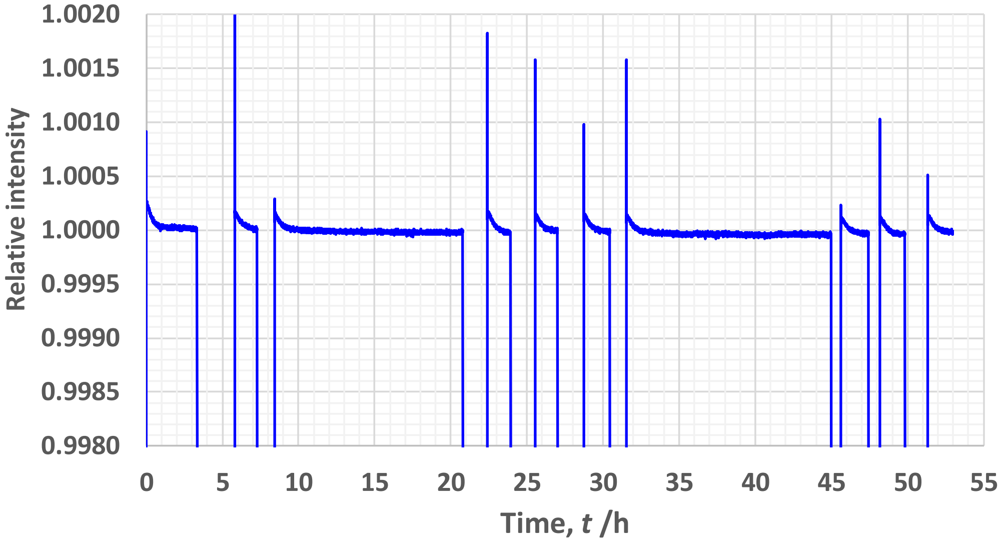


Figure 3 – Repeatability of a standard LED

The same standard LED, operating condition, and setup were also used for measuring aging rate. The standard LED was turned on and its intensity was measured every hour during most of the test time. Figure 4 shows the measurement results of the aging rate together with the room temperature for a period of more than 100 days. There is a 0.05 % initial change during the first 10 days and then no change associated with aging. During Day 96 to Day 98 the room temperature rises from 24 °C to 29 °C, which results in a rise of luminous intensity for 0.04 %. This small increase of luminous intensity may be caused by a small change of temperature of the TCM's cold plate, even though the reading of the TCM cold plate temperature does not change. Note the thermistor inside the cold plate is not thermally insulated perfectly from the ambient air and therefore it may be affected by ambient air temperature.

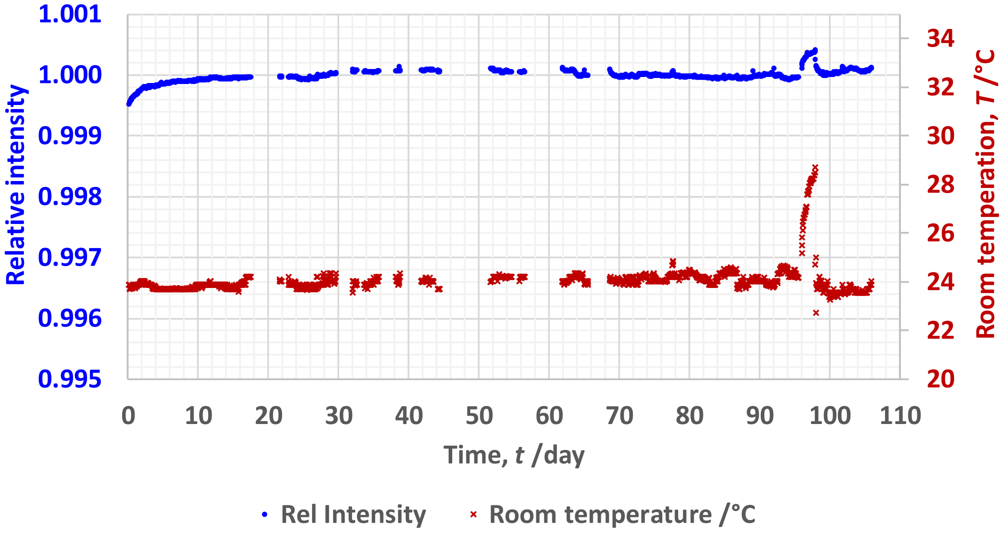


Figure 4 – Plot of aging rate

Zong, Yuqin; Zhao, Weiqiang; Miller, C Cameron. "Standard LEDS with Superior Long-Term Stability." Presented at 29th Quadrennial Session of the International Commission on Illumination (CIE), Washington, DC, US. June 14, 2019 - June 22, 2019.

Long-term stabilities of total luminous flux of 10 standard LEDs with designations of L#21 to L#30 were measured at 25 °C TCM temperature and 1 A operating current for a period of approximately three years using the NIST 2.5 m absolute integrating sphere (NIST 2018). The first measurement time is when the 10 LED packages were brand new. After the first measurement, all standard LEDs were left on with 1 A operating current and 25 °C TCM temperature (for seasoning for approximately one year) until the 2nd measurement time. After the 2nd measurement, the standard LEDs were stored in a desiccator. The measured long-term stabilities of total luminous flux of the 10 standard LEDs are shown Figure 5. Except the standard LEDs L#24 (yellow line and round dots) and L#26 (green line and round dots), variation of an LED's total luminous flux is within 0.2 % (that includes the long-term drift of the measurement system), and the variation of the average total luminous flux is within 0.1 % (the thick black line and squared dots) over the three-year period, which is superior. Further, the one-year continuous operation for seasoning resulted in virtually no change in luminous flux, which agrees with the aging test result shown in Figure 4. Therefore, the slow, time-consuming seasoning process is not needed for this type of standard LEDs operated at 1 A, and they may be used as standard LEDs from when they are brand new.

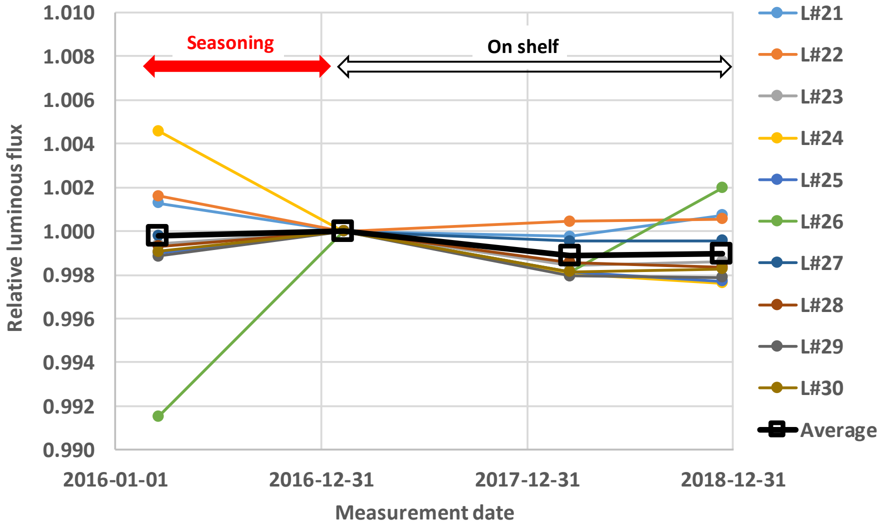


Figure 5 – Measurement result of long-term stability

4 Summary

We developed a large-chip (3 mm × 3 mm square) standard LED with virtually no aging effect. The LED has superior on-shelf long-term stability (approximately 0.1 % over three years) and can be used as a standard LED for luminous intensity, luminance, and total luminous flux.

Acknowledgements

The authors want to thank their colleagues Benjamin Tsai and Maria Nadal for their support on this research.

Reference

NIST 2018. NIST Special Publication 250-95. Yuqin Zong, Maria E. Nadal, Benjamin K. Tsai, and C. Cameron Miller. *NIST Measurement Services - Photometric Calibrations*. Gaithersburg: NIST. https://www.nist.gov/publications/nist-measurement-services-photometric-calibrations

CALIBRATION OF SPECTRORADIOMETERS USING TUNABLE LASERS

Yuqin Zong, Ping-Shine Shaw, Joseph P. Rice, C. Cameron Miller
National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA

yuqin.zong@nist.gov

Abstract

To shorten the long calibration chain when using the conventional source-based approach and therefore to reduce spectroradiometer's calibration uncertainty, we developed a detector-based approach to calibrate spectroradiometers directly against a transfer trap detector using a tunable laser. Two different detector-based methods are used for the calibration and their calibration results are in very good agreement with each other. The calibration result is also compared with that obtained using a working standard FEL lamp and the agreement is within the expanded uncertainties (k=2) of the working standard FEL lamp. This detector-based approach enables new, independent realizations of spectral irradiance or radiance scales on spectroradiometers. Such spectroradiometers can be used as instrument-based primary or transfer standards to disseminate spectral irradiance or radiance scale with a smaller uncertainty.

Keywords: Detector-based Calibration, New Method, Spectroradiometer, Tunable Laser.

1 Introduction

Spectroradiometers are typically calibrated against broadband transfer-standard sources (e.g., deuterium lamps for the deep ultraviolet (UV) region and quartz tungsten halogen lamps for the UV, visible, and infrared (IR) regions). Using this conventional source-based calibration approach, however, the uncertainty in measured spectral irradiance or radiance responsivity of a spectroradiometer is dominated by the transfer-standard sources, which is typically about 1 % in the visible and IR regions and much larger in the UV region. Such a high uncertainty arises because of a long chain of calibration in calibrating the transfer-standard sources. For example, when an FEL lamp is used for calibrating a spectroradiometer, the scale of the FEL lamp is derived and transferred in sequence from (1) primary cryogenic radiometer, (2) transfer trap detector, (3) reference field radiometer, (4) gold point blackbody, and (5) high temperature blackbody. To shorten the long calibration chain and therefore to reduce spectroradiometer's calibration uncertainty, we describe a new method for calibration of spectroradiometers directly against a transfer trap detector (i.e., bypassing most of the scale transfers mentioned above) using tunable lasers.

2 The new calibration method

The NIST automated calibration facility for calibration of detectors (Zong et al. 2012) (Woodward et al. 2018) is improved so it can also be used for calibration of spectroradiometers. A schematic for calibration of an irradiance spectroradiometer is shown in Figure 1. A fully automated tunable optical parametric oscillator (OPO) laser is used for this calibration, which has a tunable range from 210 nm to 2400 nm. The repetition rate of the OPO is one kilohertz (kHz) and its pulse width is five nanosecond (ns). The laser beam is guided into a 50 mm diameter integrating sphere using a multimode fiber so that the test spectroradiometer or standard trap detector is illuminated with a uniform beam. A laser spectrum analyser (LSA) having a spectral range of 180 nm to 1000 nm is used for measurement of the wavelength of OPO laser. The absolute wavelength standard uncertainty measured by the LSA is 5 pm and relative wavelength standard uncertainty is 1 pm. The wavelength scale of the spectroradiometer is calibrated against the LSA with a standard uncertainty on the order of a picometer (Zong 2017).

Zong, Yuqin; Shaw, Ping-Shine; Rice, Joseph P.; Miller, C Cameron. "Calibration of Spectroradiometers Using Tunable Lasers." Presented at 29th Quadrennial Session of the International Commission on Illumination (CIE), Washington, DC, US. June 14, 2019 - June 22, 2019.

The calibration is based on the measurement of the total energy of a pulsed OPO train. This avoids measurement difficulties arising from the OPO's pulse-to-pulse fluctuations (more than 10 %) and extremely low duty cycle (approximately 10^{-5}). The length of the pulsed OPO train is controlled by the laser shutter and varies from 1 s to 10 s depending on the laser power. A monitor detector is mounted near the test spectroradiometer or standard trap detector to measure the relative total energy of an OPO pulse train.

The standard trap detector and the irradiance probe of the test spectroradiometer are each mounted, in turn, to the center of the optical beam for the calibration using the substitution method. Two current integrators (also called charge amplifiers) (not shown) are used for simultaneous measurements of the total electric charge (unit: coulomb, symbol: C) from the standard trap detector and the monitor detector, respectively.

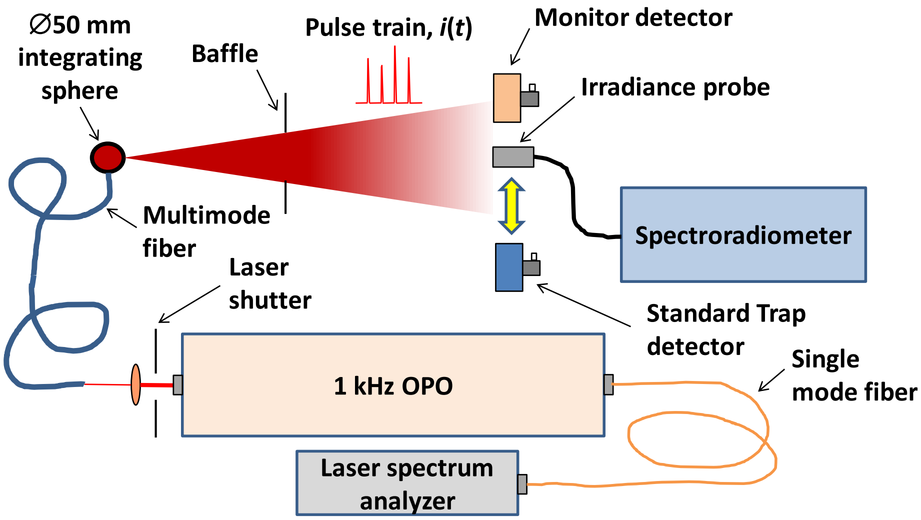


Figure 1 – Schematic for calibration of a spectroradiometer

Before the calibration, the stray light of the spectroradiometer is characterized and corrected (Zong et al. 2006) so that the spectroradiometer's response outside its bandpass is negligible. Two different methods, the slit scattering function (SSF) method and the line spread function (LSF) method, can be used for the calibration of spectroradiometers.

2.1 SSF method (or overfill method)

The SSF of an array spectroradiometer is the responsivity function of a pixel while the wavelength of the incident monochromatic source changes (Zong et al. 2006). If the SSF of a pixel is obtained by finely tuning the laser wavelength across the entire bandpass of the pixel, the sum of the response within the bandpass represents the signal when the spectroradiometer measures an imaginary broadband source with discrete wavelengths which spectrally overfills the bandpass of the pixel. Due to rapid rise-and-fall nature and narrow bandpass of an SSF, the wavelength of the laser should be tuned in a fine step (e.g., 0.1 nm) and the bandwidth of the laser should be sufficiently narrow to minimize the convolution error arising from the finite bandpass and finite bandwidth of the laser. Using this method, each pixel is considered as a filter-radiometer.

The spectral irradiance responsivity at a pixel i for an equal energy source illuminant E, $R_{\lambda,\mathrm{i}}$ (unit: count·s^{-1}/W·m^{-2}·nm^{-1}), is given by

$$R_{\lambda,\mathrm{i}} = \sum_{J \subset \mathrm{IB}} \left(\frac{y_J^{\mathrm{M}}}{Q_{J,\mathrm{trap}}^{\mathrm{M}}} \cdot R_{\mathrm{trap}}(\lambda_J) \cdot \Delta\lambda_{J,\,\mathrm{step}} \right) \tag{1}$$

where

y_J^{M} unit: count, is the signal of the pixel “i” at OPO wavelength λ_J, corrected by monitor signal and dark signal;

$Q_{J,\mathrm{trap}}^{\mathrm{M}}$ unit: C, is the electric charge of the standard trap detector at OPO wavelength λ_J, corrected by monitor signal and dark signal;

$R_{\mathrm{trap}}(\lambda_J)$ unit: A/W·m^{-2}, is the responsivity of the standard trap detector at wavelength λ_J;

$\Delta\lambda_{J,\mathrm{step}}$ unit: nm, is the wavelength interval of the scan.

As shown in Equation 1, the accuracy of wavelength interval, $\Delta\lambda_{J,\mathrm{step}}$, is critical to achieve a small calibration uncertainty because $\Delta\lambda_{J,\mathrm{step}}$ is small (e.g., 0.1 nm).

2.2 LSF method (or underfill method)

LSF of an array spectroradiometer is the description of the response at every pixel to a particular incident monochromatic light (Zong et al. 2006). If the detector array has no or negligible pixel-to-pixel spatial non-uniformity of responsivity within the bandpass of the spectroradiometer, and has no or negligible dead region between adjacent pixels, which is generally true for many charge-coupled-device (CCD) instruments, a measured LSF, at pixel j with wavelength λ_{j}, can be directly used for obtaining the spectral power responsivity of the spectroradiometer at wavelength λ_{j}, R_{j} (unit: count·s^{-1}/W·m^{-2}), by

$$R_{\mathrm{j}} = \frac{\sum_{i\subset \mathrm{IB}} y_i^{\mathrm{M}}}{Q_{\mathrm{j,trap}}^{\mathrm{M}}} \cdot R_{\mathrm{trap}}(\lambda_{\mathrm{j}}) \tag{2}$$

where

y_i^{M} unit: count, is the signal of the pixel “*i*” within the bandpass, corrected by monitor signal and dark signal;

$Q_{\mathrm{j,trap}}^{\mathrm{M}}$ unit: C, is the electric charge of the standard trap detector at OPO wavelength λ_{j}, corrected by monitor signal and dark signal;

$R_{\mathrm{trap}}(\lambda_{\mathrm{j}})$ unit: A/W·m^{-2}, is the responsivity of the standard trap detector at wavelength λ_{j}.

Using this LSF method, the spectral width of a pixel of the spectroradiometer is underfilled by the bandwidth of the laser, and the output signals of the spectroradiometer on multiple pixels within the bandpass is deconvoluted to be at 1 pixel. Note in practice, it is not required to tune the laser wavelength to be exactly on a pixel. The spectral power responsivity at a pixel can be obtained by interpolation. The spectral irradiance responsivity of the spectroradiometer at pixel j with wavelength λ_{j}, $R_{\lambda,\mathrm{j}}$ (unit: count·s^{-1}/W·m^{-2}·nm^{-1}), can be converted from the obtained spectral power responsivity, R_{j}, by

$$R_{\lambda,\mathrm{j}} = R_{\mathrm{j}} \cdot \Delta\lambda_{\mathrm{j}} \cdot C_{\mathrm{j}} \tag{3}$$

where

$\Delta\lambda_{\mathrm{j}}$ unit: nm, is the spectral width of pixel j that can be determined by measuring pixel-to-pixel wavelength interval, $\Delta\lambda_{\mathrm{p\text{-}p,j}}$, at pixel j;

C_{j} is the correction factor for non-uniformity of spectroradiometer’s spectral power responsivity on pixel j, which is the ratio of the average spectral power responsivity over the spectral width of pixel j to the spectral power responsivity at the center of pixel j.

To convert the spectral power responsivity to spectral irradiance responsivity, the spectral width, $\Delta\lambda_{\mathrm{j}}$, commonly in the range of 1 nm, must be determined with an uncertainty of picometers to achieve a small uncertainty for spectral irradiance responsivity (Zong 2017).

This LSF method does not require super fine scanning wavelength interval and super narrow bandwidth of the laser compared to the SSF method. Further, spectral irradiance responsivity at the particular wavelength can even be calibrated using a fixed wavelength laser (such as a He-Ne laser), which is very useful to check or monitor the change of the spectroradiometer in many applications.

3 Results of experimental calibration

As an example, a CCD-array spectroradiometer with the spectral range from 300 nm to 1100 nm was calibrated using the new LSF methods described above. The calibration system was first evaluated for measurement repeatability. As shown in Figure 2, the measured spectroradiometer's responsivity at 558 nm varies within 0.1 % with 5 s integration time for each data point.

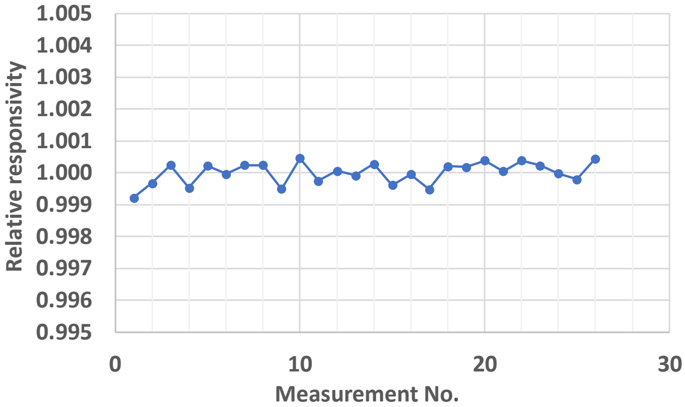


Figure 2 – Measurement repeatability of the spectroradiometer

The wavelength scale of the spectroradiometer is calibrated together with the spectral power responsivity with scanning interval of 1 nm from 350 nm 1000 nm. This limited calibration spectral range is because a trap detector may be damaged by the deep UV light and it is not stable above 1000 nm without controlling its temperature. The determined pixel-to-pixel wavelength interval is shown in Figure 3. After smoothing, the error of the spectroradiometer's pixel-to-pixel wavelength interval is reduced to a few picometers.

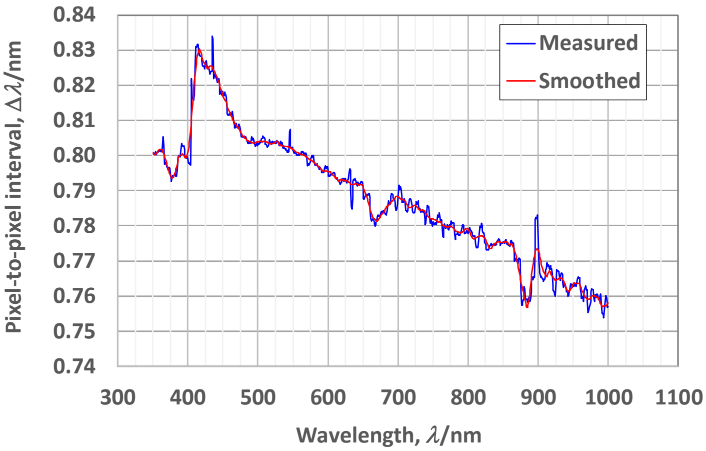


Figure 3 – Determined pixel-to-pixel wavelength interval

The spectroradiometer is also calibrated using the SSF method with a scanning interval of 0.1 nm. The spectral irradiance responsivities obtained using the SSF method are compared with those obtained using the LSF method. As shown in the Figure 4, the difference in the calibration results is approximately 0.05 %. Thus, the two methods are equivalent in terms of the calibration results for this spectroradiometer.

Zong, Yuqin; Shaw, Ping-Shine; Rice, Joseph P.; Miller, C Cameron. "Calibration of Spectroradiometers Using Tunable Lasers." Presented at 29th Quadrennial Session of the International Commission on Illumination (CIE), Washington, DC, US. June 14, 2019 - June 22, 2019.

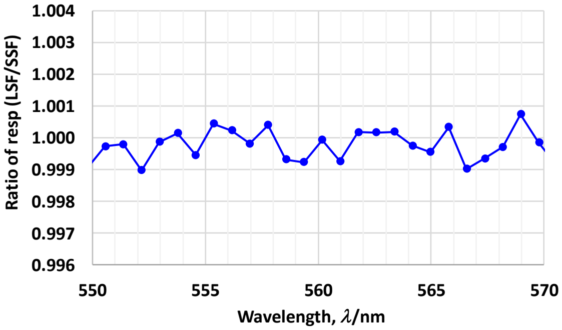


Figure 4 – Comparison of calibration results between using LSF method and using SSF method

The spectral irradiance results obtained using the detector-based LSF method are also compared to those obtained using the conventional source-based method. A spectral irradiance working standard FEL lamp is used for calibrating the spectroradiometer to obtain the spectral irradiance responsivities. Figure 5 shows the difference of the two calibration results. In the visible region from 450 nm to 700 nm, the difference does not vary much with the wavelength and it is within the expanded uncertainty (*k*=2) of the working standard FEL lamp. Outside 450 nm to 700 nm, the fluctuation of the difference is caused by a number of factors such as a low signal-to-noise ratio with the LSF method due to spectroradiometer's low responsivity and low laser power, and a large pixel-to-pixel wavelength uncertainty, etc. The glitch around 410 nm is associated with the rapid change of the pixel-to-pixel wavelength interval. The deep valley around 940 nm is mainly due to very low signal-to-noise ratio (<10:1) with the LSF method resulting from the sharp absorption band of the spectroradiometer's optical fiber.

The output laser of the OPO is its idler above 710 nm and the laser bandwidth increases rapidly with the wavelength (Zong 2012). To reduce the wavelength uncertainty above 710 nm, the OPO's signal wavelength and pump laser wavelength are measured using the LSA, and the OPO's idler wavelength is calculated using the measured signal wavelength and pump laser wavelength (as opposed to directly measurement).

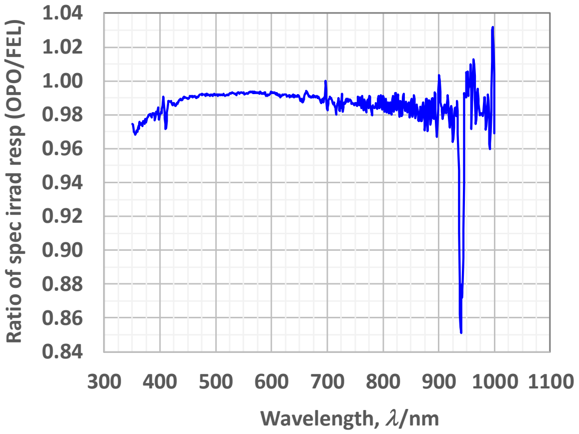


Figure 5 – Comparison of calibration results between using the detector-based LSF method and using the conventional source-based method.

4 Summary

A detector-based approach for calibrating array spectroradiometers has been developed for reducing calibration uncertainties. Two different detector-based methods, SSF method and LSF method, are used for calibrating a CCD-array spectroradiometer and their calibration results are in very good agreement with each other. The calibration result using the LSF method is also compared with that obtained using a working standard FEL lamp and the agreement is within the expanded uncertainties (k=2) of the working standard FEL lamp. The uncertainty of this detector-based calibration approach is being analysed and is expected to be much smaller than that of the conventional source-based calibration approach.

This newly developed approach enables a new, independent realization of spectral irradiance responsivity or radiance responsivity scales on spectroradiometers. Such spectroradiometers can be used as instrument-based primary or transfer standards to disseminate spectral irradiance or radiance scales with a smaller uncertainty. The detector-based calibration approach also eliminates the out-of-range stray-light error that is often the dominant source of calibration error in the UV region when a broadband standard source is used.

Acknowledgements

The authors want to thank their colleagues Steven Brown, Charles Gibson, John Woodward, Howard Yoon, Benjamin Tsai, and Maria Nadal for their support on this research.

References

NIST 2011. NIST Special Publication 250-89. Howard W. Yoon and Charles E. Gibson. *NIST Measurement Services – Spectral Irradiance Calibrations*. Gaithersburg: NIST. https://www.nist.gov/publications/sp250-spectral-irradiance-calibrations

Woodward, J. T. et al. 2018. Advances in tunable laser-based radiometric calibration applications at the National Institute of Standards and Technology, USA. *Rev. Sci. Instrum.*, 89, 091301-1 - 091301-25.

ZONG, Y. et al. 2006. Simple Spectral Stray Light Correction Method for Array Spectroradiometers. *Applied Optics*, 45, 1111-1119.

ZONG, Y. et al. 2012. A New Method for Spectral Irradiance and Radiance Responsivity Calibrations using Kilohertz Pulsed Tunable Optical Parametric Oscillators. *Metrologia*, 49, S124-S129.

ZONG, Y. 2017. Wavelength Calibration Method for Spectroradiometers with Picometer Uncertainty. *CIE*, x044:2017, 736-739.

Performance Evaluation of the NDN Data Plane Using Statistical Model Checking

Siham Khoussi[1,2], Ayoub Nouri[1], Junxiao Shi[2]
James Filliben[2], Lotfi Benmohamed[2], Abdella Battou[2], and Saddek Bensalem[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble Institute of Engineering Univ. Grenoble Alpes, VERIMAG, 38000 Grenoble, France
[2] National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA

Abstract. Named Data Networking (NDN) is an emerging internet architecture that addresses weaknesses of the Internet Protocol (IP). Since Internet users and applications have demonstrated an ever-increasing need for high speed packet forwarding, research groups have investigated different designs and implementations for fast NDN data plane forwarders and claimed they were capable of achieving high throughput rates. However, the correctness of these statements is not supported by any verification technique or formal proof. In this paper, we propose using a formal model-based approach to overcome this issue. We consider the NDN-DPDK prototype implementation of a forwarder developed at the National Institute of Standards and Technology (NIST), which leverages concurrency to enhance overall quality of service. We use our approach to improve its design and to formally demonstrate that it can achieve high throughput rates.

Keywords: NDN · SMC · Model-based design · Networking

1 Introduction

With the ever growing number of communicating devices, their intensive information usage and the increasingly critical security issues, research groups have recognized the limitations of the current Internet architecture based on the internet protocol (IP) [12]. Information-Centric Networking (ICN) is a new paradigm that transforms the Internet from a host-centric paradigm, as we know it today, to an end-to-end paradigm focusing on the content, hence more appropriate to our modern communication practices. It promises better security, mobility and scalability.

Several research projects grew out of ICN. Examples include content-centric architecture, Data Oriented Network Architecture and many others [17], but one project stood out the most and was sponsored by the National Science

The identification of any commercial product or trade name does not imply endorsement or recommendation 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.

Khoussi, Siham; Benmohamed, Lotfi; Battou, Abdella; Shi, Junxiao; Filliben, James J.; Bensalem, Saddek; Nouri, Ayoub. "Performance Evaluation of the NDN Data Plane Using Statistical Model Checking." Presented at International Symposium on Automated Technology for Verification and Analysis (ATVA2019), Taipei, TW. October 28, 2019 - October 31, 2019.

Foundation (NSF) called Named Data Networking (NDN) [19]. NDN is gaining rapidly in popularity and has even started being advertised by major networking players [1].

IP was designed to answer a different challenge, that is of creating a communication network, where packets named only communication endpoints. The NDN project proposes to generalizes this setting, such that packets can name other objects, i.e. *"NDN changes the semantics of network services from delivering the packet to a given destination address to fetching data identified by a given name. The name in an NDN packet can name anything - an endpoint, a data chunk in a movie or a book, a command to turn on some lights, etc."* [19]. This simple change has deep implications in term of routers forwarding performance since data needs to be fetched from an initially unknown location.

Being a new concept, NDN (Section 2) has not undergone any formal verification work yet. The initial phase of the project was meant to come up with proof-of-concept prototypes for the proposed architecture. This has lead to a plethora of less performing implementations in terms of packets' forwarding rates (throughput). A lot of effort was then directed to optimizing NDN forwarders' performances by trying different data structures (Hash maps) and targeting different hardware (GP-GPU). Unfortunately, validation was mainly carried using pure simulation and testing techniques.

In this work, we take a step back and try to tackle the performance problem differently. We consider a model-based approach that allows for rigorous reasoning and formal verification (Section 3). In particular, we rely on the $\mathcal{S}$BIP framework [11, 16] offering a stochastic component-based modeling formalism and Statistical Model Checking (SMC) engine. $\mathcal{S}$BIP is used along an iterative and systematic design process which consists of four phases (1) building a parameterized functional system model, which does not include performance (2) run a corresponding implementation in order to collect context information and performance measurements, characterized as probability distribution functions, (3) use these distributions to create a stochastic timed performance model and (4) use SMC to verify that the obtained model satisfies requirements of interest.

This approach is applied to verify that the NDN Data Plane Development Kit (NDN-DPDK) (an effort to develop a high performance forwarder for NDN networks at the National Institute of Standards and Technology (NIST)) can perform at high packet forwarding rates (Section 4). We investigate different design alternatives regarding concurrency (number of threads), system dimensioning (queues sizes) and deployment (mapping threads to multi-core). Using our approach, we were able to figure out what are the best design parameters to achieve higher performances (Section 5). These were taken into account by the NDN developers at NIST to enhance the ongoing design and implementation. To the best of our knowledge, this is the first work using formal methods in the context of the NDN project.

2 Named Data Networking

This section describes the NDN protocol and introduces the NDN-DPDK forwarder being designed and implemented at NIST.

2.1 Overview

NDN is a new Internet architecture different from IP. Its core design is exclusively based on naming contents rather than end points (IP addresses in the case of IP) and its routing is based on name prefix lookups [9].

The protocol supports three types of packets, namely *Interest, Data* and *Nack*. Interests are consumer requests sent to a network and Data packets are content producers replies. The Nack lets the forwarder know of the network's inability to forward Interests further. One of NDN's advantages is its ability to cache content (Data) everywhere the Data packet propagates, making the NDN router stateful. Thus, future Interests are no longer required to fetch the content from the source, instead Data could be retrieved directly from a closer node that has a cached copy.

Packets in NDN travel throughout a network as follow: first a client application sends an Interest with a name prefix that represents the requested content. Names in NDN are hierarchical (e.g., /YouTube/Alex/video1.mpg denotes a YouTube video called Video1.mpg by a Youtuber Alex). Then, this packet is forwarded by the network nodes based on its name prefix. Finally, this Interest is satisfied with Data by the original source that produced this content or by intermediate routers that cached it due to previous requests. It is also crucial to note that consecutive transmissions of Interest packets with similar name prefix might not lead to the same path each time, but could rather be forwarded along different paths each time a request is made, depending on the forwarding strategy in place. This means that the same Data could originate from different sources (producers or caches).

The NDN forwarding daemon (NFD) [3], has three different data structures: *Pending Interest Table (PIT), Content Store (CS)* and *Forwarding Interest Base (FIB).* The packet processing, according to the NDN protocol, is as follows:

1 – For Interests, the forwarder, upon receiving an Interest, starts off by querying the CS for possible copies of the Data, if a CS match is found during this operation, the cached Data is returned downstream towards the client. Otherwise, an entry is created in the PIT with its source and destination faces (communication channels that the forwarder uses for packet forwarding) for record keeping. Using the PIT, the forwarder determines whether the Interest is looped in the network by checking a global unique number called Nonce in the Interest against existing previous PIT entries. If a duplicate nonce is found the Interest is dropped and a Nack of reason *Duplicate* is sent towards the requester. Otherwise, the FIB is queried for a possible next hop to forward the Interest towards an upstream node; if there is no FIB match, the Interest is immediately dropped and replied with a Nack of reason *No Route*.

2 – For Data, the forwarder starts off by querying the PIT. If a PIT entry is found, the Data is sent to downstream nodes listed in the PIT entry, then the PIT arms a timer to signal the deletion of this entry and a copy of the Data is immediately stored in the CS for future queries. If no record is found in the PIT, the Data is considered malicious and discarded.

2.2 The NDN-DPDK Forwarder

NDN-DPDK is a forwarder developed at NIST to follow the NDN protocol and to leverage concurrency. In this paper, we evaluate its capacity to achieve high throughput using Statistical Model Checking (SMC).

The NDN-DPDK forwarder's data plane has three stages: input, forwarding, and output (Fig. 1). Each stage is implemented as one or more threads pinned to CPU cores, allocated during initialization. **Input** threads receive packets from a Network Interface Card (NIC) through faces, decode them, and dispatch them to forwarding threads. The **forwarding** thread processes Interest, Data, or Nack packets according to the NDN protocol. **Output** threads send packets via faces then queue them for transmission on their respective NIC.

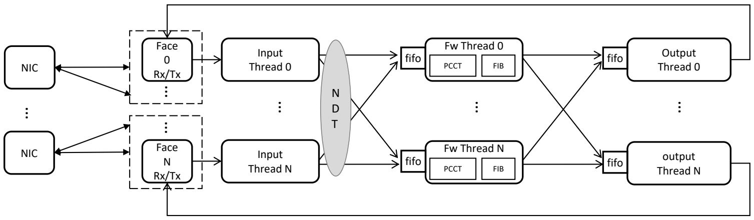


Fig. 1: Diagram of the NDN-DPDK forwarder

During forwarder initialization, each hardware NIC is provided with a large memory pool to place incoming packets. The input thread continuously polls the NIC to obtain bursts of 64 received packets. Then decodes, reassembles fragmented packets, and drops malformed ones. Then, it dispatches each packet to the responsible forwarding thread which is determined as follows: (a) For an Interest, the input thread computes SipHash of its first two name components and queries the last 16 bits of the hash value in the Name Dispatch Table (NDT), a 65,536 entry lookup table configured by the operator, to select the forwarding thread. (b) Data and Nack carry a 1-byte field in the packet header which indicates the forwarding thread that handled the corresponding Interest. Once identified, Data (or Nack) will be dispatched to the same one.

The forwarding thread receives packets dispatched by input threads through a queue. It processes each packet according to the NDN protocol, using two data structures both implemented as hash tables: (a) The FIB records where the content might be available and which forwarding strategy is responsible for the name prefix. (b) The PIT-CS Composite Table (PCCT) records which downstream node requested a piece of content, and also serves as a content cache; it combines the PIT and CS found in a traditional NDN forwarder.

The output thread retrieves outgoing packets from forwarding threads through a queue. Packets are fragmented if necessary and queued for transmission on a NIC. The NIC driver automatically frees the memory used by packets after their transmission, making it available for newly arrived packets.

3 Formal Model-based Approach

In this section, we describe the methodology used in this study which includes the underlying modeling formalism as well as the associated analysis technique.

3.1 Overview

Our methodology (Fig. 2) is based on a formal model. In order to evaluate a system's performance, its model must be faithful, i.e. it must reflect the real characteristics and behavior of this system. Moreover, to allow for exhaustive analyses, this model needs to be formally defined and the technique used for analysis needs to be trustworthy and scalable. Our approach adheres to these principles in two ways. First, by relying on the $\mathcal{S}$BIP formal framework (introduced below) that encompasses a stochastic component-based modeling formalism and an SMC engine for analysis [11]. Second, by providing a method for systematically building formal stochastic models for verification that combine accurate performance information with the functional behavior of the system.

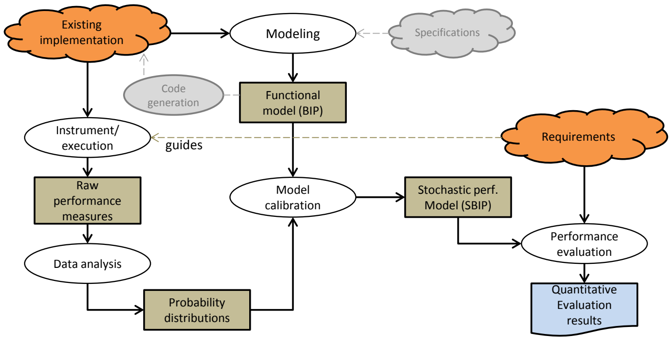


Fig. 2: Performance evaluation approach for NDN data plane.

This approach takes a functional system model and a set of requirements to verify. The functional model could be obtained from a high-level specification or an existing implementation (we use the latter in this paper). The system's implementation which could also be obtained by automatic code generation, is instrumented and used to collect performance measurements regarding the requirements of interest, e.g. throughput. These measurements are analyzed and characterized in the form of probability density functions with the help of statistical techniques such as sensitivity analysis and distribution fitting. The obtained probability density functions are then introduced in the functional model using a well defined calibration procedure [15]. The latter produces a stochastic timed model (when measurements concern time), which will be analyzed using the SMC engine.

Note that the considered models in this approach or workflow can be parameterized with respect to different aspects that we want to analyze and explore.

Basically, the defined components types are designed to be instantiated in different context, e.g. with different probability density functions thus showing different performance behaviors. While, the model considered for analysis using SMC is a specific instance for which all the parameters are fixed, some degree of parameterization is still allowed on the verified requirements.

3.2 Stochastic Component-based Modeling in BIP

BIP (Behavior, Interaction, Priority) is a highly expressive component based framework for rigorous system design [6]. It allows the construction of complex, hierarchically structured models from atomic components characterized by their behavior and their interfaces. Such components are transition systems enriched with variables. Transitions are used to move from a source to a destination location. Each time a transition is taken, component variables may be assigned new values, computed by user-defined C/C++ functions. Composition of BIP components is expressed by layered application of interactions and priorities. Interactions express synchronization constraints between actions of the composed components while priorities are used to filter among possible interactions e.g. to express scheduling policies.

The stochastic semantics of BIP were initially introduced in [14] and recently extended for real-time systems in [16]. They enable the definition of stochastic components encompassing probabilistic variables updated according to user-defined probability distributions. The underlying mathematical model behind this is a Discrete Time Markov Chain. These are modeled as classical BIP components augmented with probabilistic variables as shown in Fig. 3 and depicts a client behavior in a client-server setting where the client issues a request (**snd**) each p time units. The period p is set probabilistically by sampling a distribution function ($p \rhd$) given as a parameter of the model. Time is introduced by explicit **tick** transitions and waiting is modeled by exclusive guards on the **tick** and **snd** transitions with respect to time (captured in this example by the variable t).

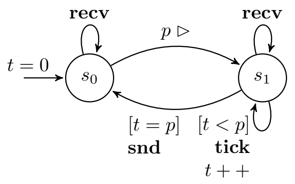


Fig. 3: A stochastic BIP component; client behavior issuing requests each time unit p.

3.3 Statistical Model Checking in a Nutshell

Statistical model-checking (*SMC*) [8, 18] is a formal verification method that combines simulation with statistical reasoning to provide quantitative answers on whether a stochastic system satisfies some requirements. It was successfully used in various domains such as biology [7], communication [4] and avionics [5]. It has the advantage to be applicable to models and implementations (provided that they meet specific assumptions) in addition to capturing rare events. The

$\mathcal{S}$BIP SMC engine [11] implements well-known statistical algorithms for stochastic systems verification, namely, Hypothesis Testing [18], Probability Estimation [8] and Rare Events. In addition, it provides an automated parameters exploration procedure. The tools take as inputs a stochastic BIP model, a Linear-time/Metric Temporal Logic (LTL/MTL) property to check and a set of confidence parameters required by the statistical test.

4 NDN-DPDK Modeling

In this section we present the modeling process of the NDN-DPDK from a functional to a stochastic timed model for throughput evaluation.

4.1 A Parameterized Functional BIP Model

Fig. 4 depicts the BIP model of the NDN-DPDK forwarder introduced in Section 2 which shows its architecture in terms of interacting BIP components that can easily be matched to the ones in Fig. 1. The presented model is parameterized with respect to the number of components, their mapping into specific CPU cores, FIFOs sizes, etc. Due to space limitation, we present in [10] the behaviors of all the components of the NDN-DPDK forwarder in Fig. 4. It is worth mentioning that the model is initially purely functional and untimed. Time is introduced later through the calibration procedure.

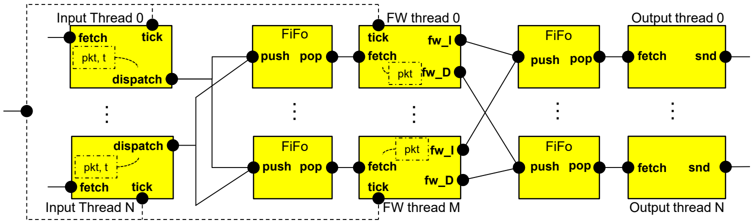


Fig. 4: A functional BIP model of the NDN-DPDK forwarder

4.2 Building the Performance Model

To build a performance model for our analysis, we consider the network topology in Fig. 5 which has a traffic generator client (consumer), a forwarder (NDN-DPDK) and a traffic generator server (producer), arranged linearly.

The green line shows the Interest packet path from the client to the producer through the forwarder and the red line indicates the Data path towards the

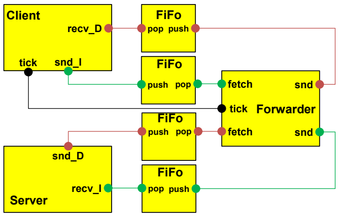


Fig. 5: Considered network topology

client. The structure of our model (Fig. 4) calls for four distribution functions to characterize performance: a) Interest dispatching latency in input threads. b) Data dispatching latency in input threads. c) Interest forwarding latency in forwarding threads. d) Data forwarding latency in forwarding threads. Notice that Nack packets are out of the scope of these experiments. We identified the following factors that can *potentially* affect the system's performance:

1. **Number of forwarding threads.** Having more forwarding threads distributes workload onto more CPU cores. The cores can compete for the shared L3 cache, and potentially increase forwarding latency of individual packets.
2. **Placement of forwarding threads onto Non Uniform Memory Access nodes (NUMA).** Input threads and their memory pools are always placed on the same NUMA node as the Ethernet adapter whereas the output threads and the forwarding threads can be moved across the two nodes. If a packet is dispatched to a forwarding thread on a different node, the forwarding latency is generally higher because memory access is crossing NUMA boundaries.
3. **Packet name length measured by the number of its components.** A longer name requires more iterations during table lookups, potentially increasing Interest forwarding latency.
4. **Data payload length.** Although the Data payloads are never copied, a higher payload length increases demand for memory bandwidth, thus potentially increasing latencies.
5. **Interest sending rate from the client.** Higher sending rate requires more memory bandwidth, thus potentially increasing latencies. It may also lead to packet loss if queues between input and forwarding threads overflow.
6. **Number of PIT entries.** Although the forwarder's PIT is a hash table that normally offers $O(1)$ lookup complexity, a large number of PIT entries inevitably leads to hash collisions, which could increase forwarding latency.
7. **Forwarding thread's queue capacity.** the queues are suspected to impact the overall throughput of the router through packet overflow and loss rates. However, it does not influence packets individual latencies.

After identifying the factors with potential influence on packet latency, we instrument the real forwarder to collect latency measurements. Then, perform statistical analysis to identify which factors are more significant. This narrows down the number of factors used and associated distribution functions.

Forwarder Instrumentation. Factors 1, 2, 3, 4, 5 and 7 can be controlled by adjusting the forwarder and traffic generator configuration, while factor 6 is a result of network traffic and is not in our control. To collect the measurement, we modified the forwarder to log packets latencies as well as the PIT size after each burst of packets. We minimized the extra work that input threads and forwarding threads have to perform to enable instrumentation, leaving the measurement collection to a separate logging thread or post-processing scripts. It is important to mention that this task does in fact introduce timing overhead. Therefore, the values obtained will have a bias (overestimate) that translates into additional latency but the trends observed remain valid.

We conducted the experiment on a Supermicro server equipped with two Intel E5-2680V2 processors, 512 GB DDR4 memory in two channels, and four Mellanox ConnectX-5 100 Gbit/s Ethernet adapters. The hardware resources are evenly divided into two NUMA nodes. To create the topology in Fig. 5, we connected two QSFP28 passive copper cables to connect the four Ethernet adapters and form two point-to-point links. All forwarders and traffic generator processes were allocated with separate hardware resources and could only communicate over Ethernet adapters.

In each experiment, the consumer transmitted either at sending intervals of one Interest per 700 ns or per 500 ns under 255 different name prefixes. There were 255 FIB entries registered in the NDN-DPDK forwarder at runtime (one for each name prefix used by the consumer), all of which pointed to the producer node. The producer would reply to every Interest with a Data packet of the same name. The forwarder's logging thread was configured to discard the first 67 108 864 samples (either latency trace or PIT size) during warm-up period, and then collect the next 16 777 216 samples and ignore the cool down session. Each experiment represents about 4 million Interest-Data exchanges. We repeated the experiment using different combinations of the factors in Table 1 and the following NUMA arrangements:

Table 1: Factors used. NUMA mapping is described below.

Factors	forwarding threads	Name length	Payload length	Sending intervals
Values	{1, 2, 3, 4, 5, 6, 7, 8}	{3, 7, 13}	{0, 300, 600, 900, 1200}	{500 ns, 700 ns}

(P1) Client and server faces and forwarding threads are all on the same NUMA,
(P2) Client face and forwarding threads on one NUMA, server face on the other,
(P3) Client face on one NUMA, forwarding threads and server face on the other,
(P4) Client face and server face on one NUMA, forwarding threads on the other.

In P1, packet latency is expected to be the smallest because all processes are placed on the same NUMA therefore, no inter-socket communication and no overhead are introduced. In P4, both Interests and Data packets are crossing NUMA boundaries twice since the forwarding threads are pinned to one NUMA whereas the client and the server faces, connected to the Ethernet adapters, reside on another. This is suspected to increase packet latency tremendously as opposed to P1, P2 and P3. These suspicions predict that placement P1 is the best case scenario and placement P4 is obviously the worst. However, we aim at getting more insight and confidence through quantitative formal analysis. This will provide a recommendation as to which placement is better suited based on the remaining parameters combinations.

Model Fitting. Before calibrating our functional BIP model with multiple distinctive probability distributions representing each combination of the factors, we choose to reduce the number of used distributions by performing a sensitivity analysis. This analysis examines the impact of several factors on the response

Khoussi, Siham; Benmohamed, Lotfi; Battou, Abdella; Shi, Junxiao; Filliben, James J.; Bensalem, Saddek; Nouri, Ayoub. "Performance Evaluation of the NDN Data Plane Using Statistical Model Checking." Presented at International Symposium on Automated Technology for Verification and Analysis (ATVA2019), Taipei, TW. October 28, 2019 - October 31, 2019.

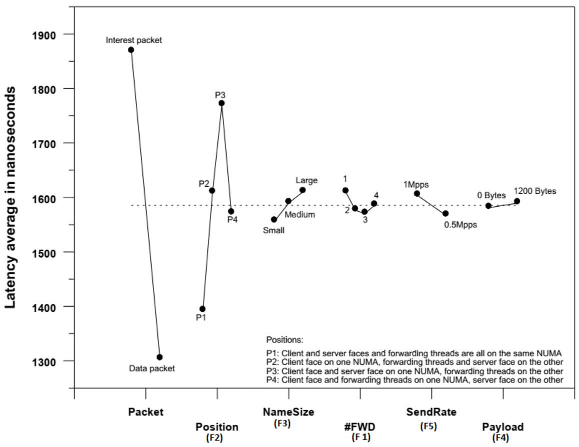


Fig. 6: Main Effects Plot for Interest and Data packets

(packet latency) and discovers the ones that are more important. In this paper, we use DataPlot [2] to produce the Main Effect Plot (Fig. 6) for factors 1 to 5.

The plot shows steeper line slopes for the packet type (packet type is not a factor. We intend to show how the NDN-DPDK forwarder processes both Interest and Data differently) as well as factors (1), (2), (3), and (5) which indicates a greater magnitude of the main effect on the latency. However, it shows almost a horizontal line for factor 4 inducing an insignificant impact on the latency. The latter is explained by the fact that the forwarder processes packet names (headers) only and doesn't read Data payloads. As for the PIT size (factor 6), it is expected to heavily increase packet latency when it is full. However, because this table's implementation is optimized for high performance and entries are continuously removed when Data packets arrive (PIT entries being satisfied), we confirmed through a correlation analysis that we can ignore this factor's impact.

Based on the analysis above, we build distribution functions for each of the factors that have greater impacts on packet latency in this study. These factors are: 1. (1) the number of forwarding threads, 2. (2) NUMA placement, 3. (3) packet name size (header), 4. (5) sending rate and, 5. (7) FIFO capacity (FIFO impacts the loss rates and not individual packet latency). We refer the reader to [10] to understand how we obtained the probability distributions for these factors.

Model Calibration. Calibration is a well defined model transformation that transforms functional components into stochastic timed ones [13]. In this section, we use the probability distributions obtained above to calibrate the functional BIP model of the NDN-DPDK forwarder shown in Fig. 4. Due to space limitations, we refer the reader to [10] where we describe the calibrated models of all the BIP components of the NDN-DPDK forwarder.

In the next section, we perform SMC on the calibrated model of the NDN-DPDK forwarder and explain the results.

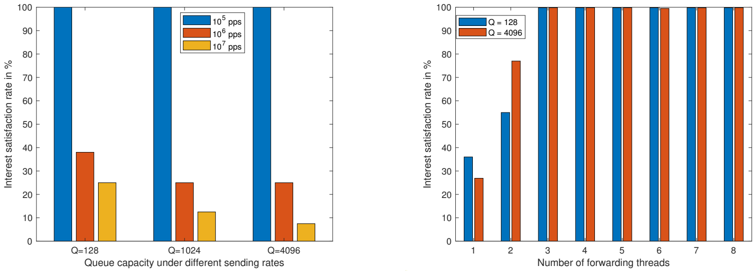
5 Performance Analysis using SMC

5.1 Experimental Settings

We run the SMC tests using the probability estimation algorithm (PE) with a required confidence of $\alpha = 0.1$ and a precision of $\delta = 0.1$. Each test is configured with a different combination of values for the factors previously presented. And each execution of a test with a single set of parameters generates a single trace. The property evaluated with the SMC engine is: *Estimate the probability that all the issued Interests are satisfied, i.e. a Data is obtained in return for each Interest*. The SMC result is a probability estimation $\hat{p}$ which should be interpreted as being within the confidence interval $[\hat{p} - \delta, \hat{p} + \delta]$ with probability at least $(1 - \alpha)$. In the experiments below, the shown results corresponds to $\hat{p} = 1$.

5.2 Analyses Results

Queues Dimensioning. First, we explore the impact of sizing forwarding threads queues. Each forwarding thread has an input queue. Initially, we consider a model with a single forwarding thread and vary its queue capacity with 128, 1024 or 4096 (in packets). Then set the client's sending rate to: 10^5 packets per second (pps), 10^6 pps or 10^7 pps. The results are shown in Fig. 7a. The Y-axis represents the Interest satisfaction rate such that 100 % (resp. 0 %) indicates no loss (resp. 100 % loss) and the x axis represents the queue capacity under different sending rates.



(a) One Forwarding thread with different sending rates.

(b) Many Forwarding threads with a sending rate set to 10^6 pps.

Fig. 7: Exploration results of the Forwarding threads queues sizes.

Fig. 7a indicates that at 10^5 pps (blue), the Interest satisfaction rate is 100%. This means that the forwarder (with one forwarding thread) is capable of handling all packets at this sending rate (10^5 pps of packet size 1500 bytes is equivalent to 1.2 Gbps), under any queue size. However, under a faster sender rate (where a single forwarder shows signs of packet loss) we unexpectedly observed a better Interest satisfaction rate with a smaller queue (Q=128). After a thorough investigation of the real implementation, we found out that the queues don't have proper management in terms of insertion and eviction policies that would give priority to Data over Interest packets. In the absence of such policy, more Interests would be queued while Data packets would be dropped resulting in Interests not being satisfied, thus lower performance (Interest satisfaction rate). *It is thus advised for the final implementation of the NDN-DPDK forwarder, to use a queue capacity smaller than* 128 *packets when the forwarder has a single forwarding thread and packets are sent at a fast rate.*

Similarly, we explore whether this observation remains true with more forwarding threads. In order to do that, we run SMC again on eight different models each with a different number of forwarding threads (1 to 8) under a sending rate of 10^6 pps (1 Interest per 1 us) where a loss rate was observed in Fig. 7a. Then, we experimented with two queue capacities, namely 128 and 4096 packets. The results are reported in Fig. 7b. The x Axis represents the number of forwarding threads while the y axis depicts the Interest satisfaction rate.

We observe that the queue size matters mainly in the case of a model with one and two forwarding threads. In fact, for a two threads model, a bigger queue size is preferred to maximize the performance, unlike when a single thread is used. As for the other six models, both sizes achieve almost 100 % Interest satisfaction. This is due to the fact that three forwarding threads or more are capable of splitting the workload at 10^6 pps and can pull enough packets from each queue with a minimum loss rate of 0.02 % . *This result stresses that, to avoid being concerned about a proper queue size, more threads are needed for handling a faster sending rate with minimum Interest loss.*

NUMA placement, number of forwarding threads and packet name length. Another aspect to explore, is the impact of mapping the forwarding threads and/or NDN Faces to the two NUMA nodes (0, 1) under different sending rates and for multiple name lengths where Face 0 exchanges packets with the client and Face 1 with the server. To do that, we consider the four NUMA arrangements (P1), (P2), (P3) and (P4) in section 4 as well as the factors in Table 1 in the SMC analysis.

In Figs. 8 to 13, each row represents experiments with similar packet name lengths {small=3, medium=7, large=13} and a queue capacity of 4096. The right-hand column indicates results for a faster sending rate of $2 * 10^6$ pps (500 ns interval) while the left-hand one shows results for a slower sending rate of $1.42 * 10^6$ pps (700 ns interval). The six figures includes four curves where each corresponds to the four NUMA arrangement options: P1 to P4.

The six Figs. 8 to 13 show that Interest satisfaction rates scale up with the increase of forwarding threads then reach a saturation plateau where adding more threads can no longer improve the performances. Furthermore, with fewer forwarding threads, the loss rate is unavoidable and exceeds 80 %. This is because the sending rate is faster than the forwarding threads processing capabilities

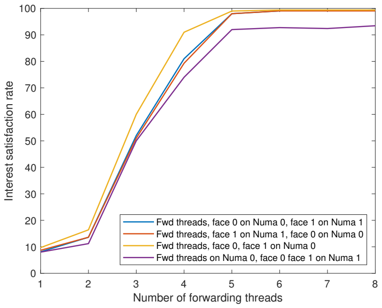


Fig. 8: small names, 700 ns

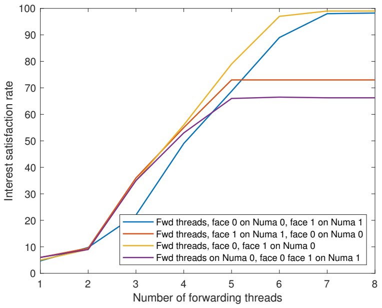


Fig. 9: small names, 500 ns

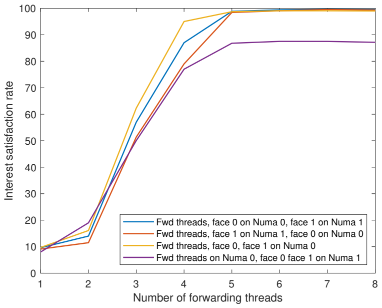


Fig. 10: medium names, 700 ns

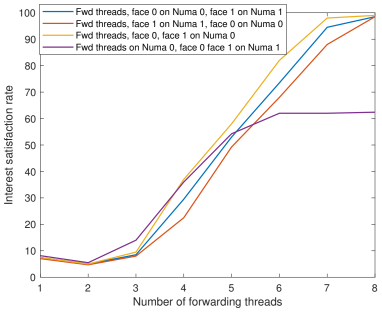


Fig. 11: medium names, 500 ns

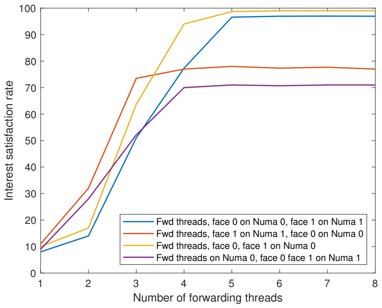


Fig. 12: large names, 700 ns

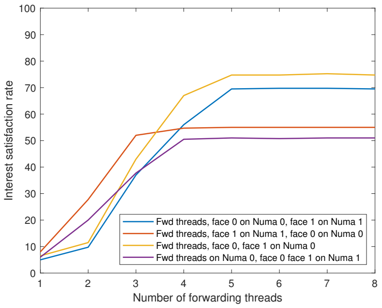


Fig. 13: large names, 500 ns

causing their FIFO queues to saturate and start dropping packets frequently. However, under a slower sending rate and packets with small, medium and large name lengths (3, 7, 13), Figs. 8, 10 and 12 show that a maximum satisfaction rate of over 90 % is achievable with only five forwarding threads. Whereas when the client is generating packets faster at 2 Mpps, a saturation plateau of over 90 % is reached at six threads or more for small and medium names (Figs. 9 and 11) and a plateau of slightly over 70 %, with five threads, for larger names (Fig. 13). Also, Figs. 8 and 10 demonstrate that placing all processes (threads and faces) on a single NUMA (placement P1) outperforms the other three options. This observation is explained by the absence of inter-socket communication thus less timing overhead added such as in the case of the purple plot where packets are crossing NUMA boundaries twice from Face 0 to the forwarding threads then through Face 1 and back (placement P4).

Figs. 9 and 11 show the impact of increasing the sending rate on packets with smaller names. In this case, it is preferred to also position all the processes on one NUMA such as the case of the yellow plot of the P1 series because NUMA boundary crossing usually downgrades the performance. In fact, the difference between no NUMA crossing and the double crossing (yellow and purple series respectively) is approximately 30 % loss rate with more than five threads. The second best option P2 which is placing the forwarding threads on the NUMA receiving Interest packets with Face 0 (NUMA hosting the Ethernet adapter that receives Interests from the Client). However, when the number of threads is not in the saturation zone and the threads get overworked and start to loose packets, it is recommended to opt for placement P3. *Based on these results, we recommend that for small to medium names, to use a maximum of eight threads but no less than five arranged as in placement P1 for optimum performances under a slower or a faster sending rate.*

With a larger name however, Fig. 12 depicts an unexpected behaviour when using three threads or less. In this case, placing the forwarding threads on the same NUMA as Face 1 (which is the Ethernet adapter connected to the server and receives Data packets), surpasses the other three options. Our explanation is that since forwarding threads take longer times to process incoming packets due to their longer name and timely lookup, particularly for Interests as they are searched by names inside the two tables (PCCT and FIB) rather than a token such as the case for Data packets. Placing the forwarding threads with the Data receiving Ethernet adapter connected to Face 1, has the potential to yield better results by quickly processing packets after a quick token search especially when the workload is bigger than the threads' processing capacity. When the sending rate is increased, the same results are observed in Fig. 13 for a similar name length but with a decrease in performance. *Thus, we recommend for larger names to use NUMA arrangement P3 only when the number of forwarding threads is less than three regardless of the sending rate (not advised due to high loss rate).*

6 Lessons learned and future work

This study shed light on a new networking technology called Named Data Networking (NDN) and its forwarding daemon. Ongoing NDN research includes the development of high-speed data plane forwarders that can operate at a hundred gigabits per second while using modern multi-processor server hardware and kernel bypass libraries. In this paper, we discussed the results of a performance evaluation effort we undertook to reach well-founded conclusions on how the NDN forwarder prototype developed by NIST (NDN-DPDK) behaves in a network in terms of achievable Interest satisfaction rate.

We conducted an extensive analysis under different factors such as the number of threads carrying tasks and function mapping to CPUs, using a model-based approach and statistical model checking. Given the wide array of design parameters involved, this effort contributes valuable insights into protocol operation and guides the choice of such parameters. The use of statistical model checking for performance analysis allowed us to discover potential sub-optimal operation and propose appropriate enhancement to the queue management solution. This has been taken into account in the ongoing NDN-DPDK forwarder implementation. Moreover, our extensive analysis provides a characterization of the achievable forwarding throughput for a given forwarder design and available hardware resources which would not have been possible to obtain, with such controllable accuracy, using traditional measurements and statistic methods. Furthermore, these results were communicated and shared with members of the NDN community in a conference throughout a poster interaction and gained attention from researchers who were interested in the methodology and its applications. In addition to that, the use of a BIP model refined at the right level of abstraction allows the generation of executable code that could be used instead of the real implementation.

It is important to note however, that our analysis depends largely on a stochastic model obtained using samples of data collected from the actual implementation of the forwarder which is suspected to have introduced timing overhead. Nevertheless, the trends observed throughout this study remain accurate and have provided valuable insight to the actual code. In the future, this analysis will be extended to answer the reverse question, namely **Given a desired throughput, what is the best hardware setup and the forwarder design to use?** Rather than the question **Given a hardware setup and a forwarder design, what is the maximum achievable throughput?** that we have investigated in this paper.

References

1. Brown, b. (2019). cisco, ucla & more launch named data networking consortium. [online] network world., `https://www.networkworld.com/article/2602109/ucla-cisco-more-join-forces-to-replace-tcpip.html`

2. Dataplot homepage, `https://www.itl.nist.gov/div898/software/dataplot/homepage.htm`
3. NFD Developer's Guide. Tech. rep., `http://named-data.net/techreports.html`
4. Basu, A., Bensalem, S., Bozga, M., Caillaud, B., Delahaye, B., Legay, A.: Statistical Abstraction and Model-Checking of Large Heterogeneous Systems. In: Forum for fundamental research on theory, FORTE'10. LNCS, vol. 6117, pp. 32–46. Springer
5. Basu, A., Bensalem, S., Bozga, M., Delahaye, B., Legay, A., Siffakis, E.: Verification of an AFDX infrastructure using simulations and probabilities. In: Runtime Verification, RV'10. LNCS, vol. 6418. Springer (2010)
6. Basu, A., Bozga, M., Sifakis, J.: Modeling heterogeneous real-time components in bip. In: Proceedings of the Fourth IEEE International Conference on Software Engineering and Formal Methods. pp. 3–12. SEFM'06, IEEE Computer Society, Washington, DC, USA (2006)
7. David, A., Larsen, K.G., Legay, A., Mikucionis, M., Poulsen, D.B., Sedwards, S.: Statistical model checking for biological systems. Int. J. Softw. Tools Technol. Transf. **17**(3), 351–367 (Jun 2015)
8. Hérault, T., Lassaigne, R., Magniette, F., Peyronnet, S.: Approximate Probabilistic Model Checking. In: International Conference on Verification, Model Checking, and Abstract Interpretation, VMCAI'04. pp. 73–84 (January 2004)
9. Jacobson, V., Smetters, D.K., Thornton, J.D., Plass, M.F., Briggs, N.H., Braynard, R.L.: Networking Named Content (2009), `https://named-data.net/wp-content/uploads/Jacob.pdf`
10. Khoussi, S., Nouri, A., Shi, J., Filliben, J., Benmohamed, L., Battou, A., Bensalem, S.: Performance evaluation of a NDN forwarder using statistical model checking. CoRR **abs/1905.01607** (2019), `http://arxiv.org/abs/1905.01607`
11. Mediouni, B.L., Nouri, A., Bozga, M., Dellabani, M., Legay, A., Bensalem, S.: $\mathcal{S}$BIP 2.0: Statistical model checking stochastic real-time systems. In: Automated Technology for Verification and Analysis - 16th International Symposium, ATVA 2018, Los Angeles, CA, USA, October 7-10, 2018, Proceedings. pp. 536–542 (2018)
12. Named data networking project. Tech. rep., USA (Oct 2010), `http://named-data.net/techreport/TR001ndn-proj.pdf`
13. Nouri, A.: Rigorous System-level Modeling and Performance Evaluation for Embedded System Design. Ph.D. thesis, Grenoble Alpes University, France (2015)
14. Nouri, A., Bensalem, S., Bozga, M., Delahaye, B., Jegourel, C., Legay, A.: Statistical model checking QoS properties of systems with SBIP. Int. J. Softw. Tools Technol. Transf. (STTT) **17**(2), 171–185 (April 2015)
15. Nouri, A., Bozga, M., Molnos, A., Legay, A., Bensalem, S.: *ASTROLABE*: A rigorous approach for system-level performance modeling and analysis. ACM Trans. Embedded Comput. Syst. **15**(2), 31:1–31:26 (2016)
16. Nouri, A., Mediouni, B.L., Bozga, M., Combaz, J., Bensalem, S., Legay, A.: Performance evaluation of stochastic real-time systems with the sbip framework. International Journal of Critical Computer-Based Systems **8**(3-4), 340–370 (2018)
17. Xylomenos, G., Ververidis, C.N., Siris, V.A., Fotiou, N., Tsilopoulos, C., Vasilakos, X., Katsaros, K.V., Polyzos, G.C.: A survey of information-centric networking research. IEEE Communications Surveys Tutorials **16**(2), 1024–1049 (2014)
18. Younes, H.L.S.: Verification and Planning for Stochastic Processes with Asynchronous Events. Ph.D. thesis, Carnegie Mellon (2005)
19. Zhang, L., Afanasyev, A., Burke, J., Jacobson, V., claffy, k., Crowley, P., Papadopoulos, C., Wang, L., Zhang, B.: Named data networking. SIGCOMM Comput. Commun. Rev. **44**(3), 66–73 (Jul 2014)

Contemporary Mathematics

Complex Variables, Mesh Generation, and 3D Web Graphics: Research and Technology Behind the Visualizations in the NIST Digital Library of Mathematical Functions

Bonita V. Saunders

Abstract. In 2010, the National Institute of Standards and Technology (NIST) launched the Digital Library of Mathematical Functions (DLMF), a free online resource containing definitions, recurrence relations, differential equations, and other crucial information about mathematical functions useful to researchers working in application areas in the mathematical and physical sciences. Although the DLMF was designed to replace the widely cited National Bureau of Standards(NBS) Handbook of Mathematical Functions commonly known as Abramowitz and Stegun (A&S), the goal was a compendium far beyond a book on the web, incorporating web tools and technologies for accessing, rendering, and searching math and graphics content. This paper focuses primarily on the research and technical challenges involved in creating the DLMF's graphics content, and in particular, its interactive 3D visualizations, where users can explore more than 200 graphs of high level mathematical function surfaces.

1. Introduction

In 2010, after a multi-year effort dating back to the late 1990s, the National Institute of Standards and Technology (NIST) released the Digital Library of Mathematical Functions (DLMF)[**9**], a free online resource containing definitions, recurrence relations, differential equations, and other crucial information to aid in the understanding and computation of mathematical functions that arise in application areas in the mathematical and physical sciences. Although the DLMF might be viewed as an update and replacement for the 1964 Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (A&S) [**1**], in reality, it is quite different in both focus and content.

A&S was originally known for its tables. Its existence can be traced back to the Mathematical Tables Project created in 1938 by NIST's predecessor, the National Bureau of Standards (NBS), to address a crucial need for accurate tables to assist in the computation of functions commonly occuring in practical problems [**5**]. The project was administered by the Works Projects Administration, a New Deal

2010 *Mathematics Subject Classification.* Primary 65D17.

Saunders, Bonita. "Complex Variables, Mesh Generation, and 3D Web Graphics: Research and Technology Behind the Visualizations in the NIST Digital Library of Mathematical Functions." Presented at Joint Mathematics Meetings 2019, Baltimore, MD, US. January 16, 2019 - January 19, 2019.

agency of President Franklin Roosevelt. Highly educated, but out of work mathematicians and physicists supervised a staff of human 'computers' who performed calculations for reference tables of function values. From 1938 to 1946, 37 volumes of tables were published, including tables of trigonometric functions, logarithms, the exponential function, and probability functions [**5**]. Realizing the importance of having the information all in one place, NBS mathematician Milton Abramowitz, a technical leader of the Mathematical Tables Project, eventually pushed for the publication of a compendium of tables and related material. This compendium, with emphasis on higher level functions such as Bessel functions, hypergeometric functions, and elliptic functions was published as the Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables in 1964. It is often simply called Abramowitz & Stegun or A&S in honor of its editors Milton Abramowitz and Irene Stegun who were Chief and Assistant Chief, respectively, of the Computation Laboratory of the NBS Applied Mathematics Division around the start of the project in 1956. Stegun took over the project and shepherded it to completion when Abramowitz died suddenly of a heart attack in 1958[**5**].

A casual glance at A&S clearly shows that tables predominate the handbook, but each function also includes related material such as formulas representing differential equations, definite and indefinite integrals, inequalities, recurrence relations, power series, asymptotic expansions, polynomial and rational approximations, graphs, and other qualitative information that might be useful for understanding and computing values of the function. For practicioners this "related material" has moved to the forefront over the years, while the tables have receded in importance due to the prevalence of reliable numerical software and computer algebra packages that have severely decreased the need for tables for computing function values by interpolation. Acknowledging this, it was decided that tables would not be included in the design for the DLMF. Nevertheless, the addition of new chapters on functions of growing significance and information on new properties of existing A&S functions increased the DLMF's content significantly over that of A&S.

The fact that the DLMF is web-based has opened up many possibilities that are still being explored. Navigational tools and hyperlinks allow the user to move around the site in a variety of ways. A database of metadata provides information for pop-up boxes where users can find links to sources, defined variables, cross references, alternative text formats such as LaTeX, MathML, and image formats, or notes on changes made to content. With the DLMF's math-aware search engine users can search by function names, particular formulas, and in some cases types of functions, for example, 'trig'. The site includes 600 2D and 3D plots along with 200 dynamic interactive visualizations where users can explore the graphs of elementary and high level mathematical functions. An overview of the technological capabilities in the DLMF was recently published in a Physics Today article[**20**]. In this paper we take a more in-depth view of the DLMF's graphics by looking at the ongoing research and development behind its interactive 3D visualizations.

Section 2 describes the techniques used to construct the computational grids for plotting the graphs shown in the visualizations. In Section 3 we look at our utilization and advancement of techniques for displaying interactive 3D graphics on the web and discuss some of the interesting capabilities available in the DLMF visualizations. Section 4 discusses ongoing challenges and future areas for research.

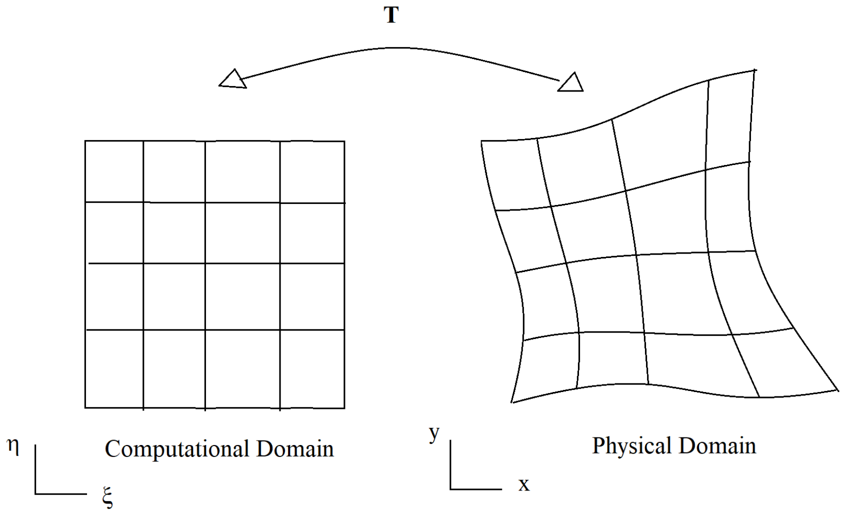


Figure 1. Numerical grid generation is defined by a curvilinear coordinate map from a canonical domain to the oddly shaped physical domain prescribed by the application.

2. Grid Generation

During the design phase of the DLMF in the late 1990s and early 2000s little thought was given to how one might plot and render complex function graphs on the web. The focus was on the primary mathematical content that chapter authors would be asked to write. However, once a draft of the first chapter, on Airy and related functions, was written, we began looking at the best way to create the illustrative graphs the chapter needed. On looking at computer algebra systems we found that most did an excellent job plotting 2D graphs, which could be exported to an acceptable format for viewing on the web. For 3D graphs of function surfaces, the story was quite different. Most systems had rudimentary or non-existent machinery for properly clipping the graph when it was necessary to restrict the displayed range to illuminate significant features, such as poles or zeros. In one case the surface was properly clipped when viewed inside the system, but the unclipped data reappeared when the plot data was exported to a file. The problems we observed led us to design our own software for grid generation and, as we show in the next section, inspired our development of visualizations utilizing emerging web 3D technology.

We solved the clipping problem by computing the function over a 2D grid whose boundary included a level curve, or contour, of the function. We created the grid by using numerical grid generation, which defines a curvilinear coordinate system through a map from a canonical domain, such as a square in 2D, to the physical domain of interest, as shown in Figure 1.

Numerical grid generation is just one of several methods for creating a grid, or mesh, for solving problems over an oddly shaped domain. It has often been used with finite difference methods to solve partial differential equations (PDEs) governing flow around interesting geometries such as an airplane wing, ship's hull, or automobile body. Numerical grid generation is sometimes called structured grid generation because of the natural array order of the grid points on the physical domain [**23**]. Unstructured methods such as Delaunay triangulations, quadtree methods, or hybrid methods that combine both structured and unstructured meshes are often the methods of choice for extremely complex geometries. However, they require the storage of grid node connectivity information that can sometimes cause memory issues.

Structured methods can require a bit of ingenuity if the boundary shape is complex, but they may allow one to write more efficient code for some applications. In our case the structured order facilitated the coding of the interactive features for our visualizations. Our code is based on an algorithm we initially designed for problems in aerodynamics and solidification theory [**13, 14, 16**]. We modified the code to accurately approximate function boundary and contour data as well as capture significant function attributes such as poles, zeros, branch cuts, and other singularities.

When numerical grid generation is used for solving PDEs, the coordinate mapping must be one-to-one and onto to ensure invertibility. The goal is to transform the equations from the physical domain to equations over a simpler canonical domain where the difference equations and boundary conditions are easier to apply. Although, in our case, the 2D grid over the physical domain becomes the computational grid for the function being plotted, the same one-to-one correspondence is still needed since it ultimately affects the accuracy of the surface clipping and the smoothness of the colormap when the surface is rendered on the web [**15, 17**]

Our basic algorithm constructs a curvilinear coordinate spline mapping $\mathbf{T}$ from the unit square I_2 to the physical domain and is defined by

$$\mathbf{T}(\xi,\eta) = \begin{pmatrix} x(\xi,\eta) \\ y(\xi,\eta) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{ij}B_{ij}(\xi,\eta) \\ \sum_{i=1}^{m}\sum_{j=1}^{n}\beta_{ij}B_{ij}(\xi,\eta) \end{pmatrix} \tag{2.1}$$

where each B_{ij} is the tensor product of cubic B-splines. Therefore, $B_{ij}(\xi,\eta) = B_i(\xi)B_j(\eta)$ where B_i and B_j are elements of cubic B-spline sequences associated with finite nondecreasing knot sequences, say, $\{p_i\}_1^{m+4}$ and $\{q_j\}_1^{n+4}$, respectively [**13**].

To quickly obtain initial α_{ij} and β_{ij} coefficients for $\mathbf{T}$ we construct a transfinite blending function mapping [**11, 23, 10**] that interpolates the boundary of the physical domain. Conveniently, the spline coefficients can be divided into boundary coefficients that map the boundary of the square onto the boundary of the physical domain, and interior coefficients [**13, 14**], which hopefully map the interior of the square onto the interior of the physical domain. Their initial values are obtained by evaluating the transfinite interpolant at knot averages as described in [**4**], to produce a shape preserving approximation that reproduces straight lines and preserves convexity. If more accuracy is needed on part of the boundary, we use de Boor's SPLINT routine [**4**] to find coefficients that produce a cubic spline interpolant of that side. It is important that the spline knots and boundary coefficients be chosen carefully to produce an accurate representation of the physical boundary.

For simple boundaries, the initial coefficients produce a grid that is adequate for most applications, but if the boundary is more complicated or highly nonconvex, modifications of the coefficients are often necessary. To improve the grid, we fix

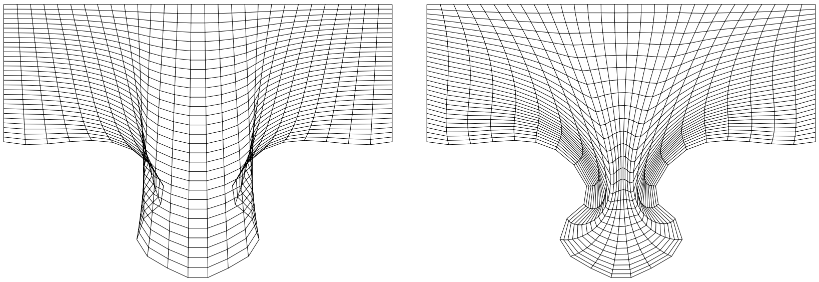


FIGURE 2. Initial and optimized puzzle grids.

the boundary coefficients and use a variational method to find interior coefficients that minimize the functional

$$F = \int_{I_2} \left(w_1 \left\{ \left(\frac{\partial J}{\partial \xi} \right)^2 + \left(\frac{\partial J}{\partial \eta} \right)^2 \right\} + w_2 \left\{ \frac{\partial \mathbf{T}}{\partial \xi} \cdot \frac{\partial \mathbf{T}}{\partial \eta} \right\}^2 \right) dA \tag{2.2}$$

where $\mathbf{T}$ denotes the grid generation mapping, J is the Jacobian of the mapping, and w_1 and w_2 are weight constants. This integral controls mesh smoothness and orthogonality. A large value for w_1 will decrease the variance in Jacobian values at nearby points, making the grid smoother. The w_2 weighted term represents the dot product of the tangent vectors $\partial\mathbf{T}/\partial\xi$ and $\partial\mathbf{T}/\partial\eta$. Therefore, minimizing this term enhances grid orthogonality. A change of variables shows this term to be equivalent to the volume weighted version of the orthogonality term in the Brackbill and Saltzman functional [**6**]. Figure 2 shows the initial and optimized grids for a physical domain shaped like a puzzle piece.

Figure 3 shows the computational grid and Riemann zeta function surface plot created using it. The grid boundaries, including the exterior boundary and the interior one around the pole, contain contour data for a height of 3. Computing the function over the grid produces a smooth clipping of the surface. A number of the "non-trivial" zeros of the Riemann zeta function can be viewed by exploring the figure on the DLMF site [**9**]. In our original code we input the location of the zeros to guarantee that there are gridpoints there. We are currently working on an algorithm that will automatically move gridpoints to the vicinity of a zero.

The current algorithm contains two significant changes over the original. First, we have added an adaptive term $w_3\{uJ^2\}$ to the integrand of the functional, where w_3 is a weight constant, and u represents external criteria for adapting the grid. If we were solving a system of partial differential equations, u might represent the gradient of the evolving solution or an approximation of truncation error. For our purposes, we want u to contain curvature and gradient information related to the function surface. The goal is to create a grid generation system that adaptively moves gridpoints to areas of high curvature or large gradient. With a change of variables this term is equivalent to the weighted volume variation, or adaptive, component of the Brackbill and Saltzman functional [**6, 23**]. Therefore, the enhanced integral should allow some control over mesh smoothness, orthogonality, and through u, permit an adaptive concentration of grid lines.

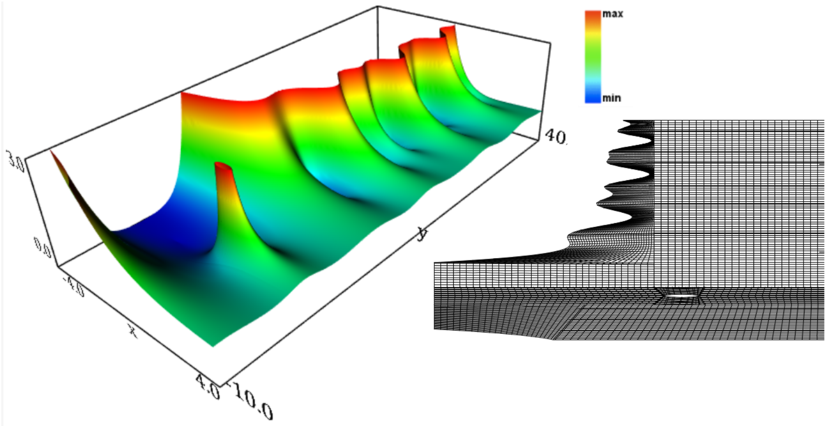


Figure 3. Riemann zeta function surface obtained by computing function over grid shown.

Second, a more fundamental change in our algorithm is replacing the mapping $\mathbf{T}$ by a composite mapping $\mathbf{T}^* = \mathbf{T} \circ \mathbf{\Phi}$ where $\mathbf{\Phi}$ is a tensor product spline mapping from the unit square I_2 to I_2 with its own coefficients and knot sequences [**18**]. Adaptive methods typically construct a reference grid [**22**] or distribution mesh [**21, 13**] by moving points on the canonical domain based on some adaptive criteria. The reference/distribution mesh is then mapped to the physical domain to create an adaptive mesh there. Mathematically this could be viewed as the composition of two maps where one maps the canonical domain to itself and the other maps the canonical domain to the physical domain. DeRose, et al., created composite maps in B-spline or Bézier form [**7, 8**].

For now, we have not tried to create a simple representation for our composite map, that is, we leave $\mathbf{\Phi}$ and $\mathbf{T}$ in their separate forms. The boundary coefficients of $\mathbf{T}$ can remain fixed while the coefficients of $\mathbf{\Phi}$ are adjusted to reparameterize the boundary points. The $\mathbf{\Phi}$ map can be used to create a reference grid that produces the desired adaptive effect without disturbing the accuracy of the boundary approximation.

After choosing initial $\mathbf{T}$ and $\mathbf{\Phi}$ coefficients that approximate transfinite interpolation we can improve our final physical grid, that is, the smoothness, orthogonality, or concentration of grid points by minimizing the following functional with respect to either the $\mathbf{\Phi}$ coefficients or interior $\mathbf{T}$ coefficients:

(2.3)
$$F^* = \int_{I_2} \left(w_1 \left\{ \left(\frac{\partial J^*}{\partial \xi} \right)^2 + \left(\frac{\partial J^*}{\partial \eta} \right)^2 \right\} + w_2 \left\{ \frac{\partial \mathbf{T}^*}{\partial \xi} \cdot \frac{\partial \mathbf{T}^*}{\partial \eta} \right\}^2 + w_3 \{ u {J^*}^2 \} \right) dA$$

where $*$ has been added to indicate the terms are associated with the composite mapping $\mathbf{T}^*$. Then J^*, the Jacobian of $\mathbf{T}^*$ is the product of J and J_Φ where J is the Jacobian of $\mathbf{T}$ and J_Φ, the Jacobian of $\mathbf{\Phi}$. To simplify our notation, $*$ is not added to the weight constants or u. Figure 4 shows a puzzle shaped grid adapted to the vertical line $x = 5.5$. We first optimize with respect to the interior $\mathbf{T}$ coefficients

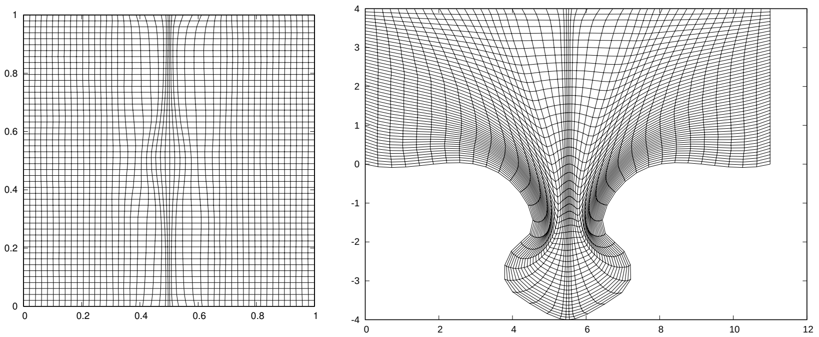


Figure 4. Reference grid and adapted puzzle grid.

to obtain an acceptable physical grid as shown in Figure 2. We then optimize with respect to the $\mathbf{\Phi}$ coefficients to adapt the grid to the line. Our adaptive function u is defined by

$$u(x, y) = e^{-50(x-5.5)^2}. \tag{2.4}$$

We have defined other expressions for u to adapt grids to circular arcs, circles, and intersecting lines [**18**]. We are now focusing on improving the performance of the code and experimenting with various definitions of u to capture function curvature and gradient data. Also, since our function data is computed using a variety of codes and computer algebra packages, we are also working on the integration of our grid generation code with various software packages and systems.

3. 3D Web Graphics and DLMF Implementation

The individual chapters of both A&S and the DLMF were authored by various mathematicians and physicists of note. One A&S author was Philip J. Davis, who was at NBS at that time. Davis prepared the chapter on gamma and related functions, which he designed to serve as a model for the other authors. Davis hired Frank W.J. Olver who authored the A&S chapter on Bessel functions. Years later, Olver would serve as DLMF Mathematics Editor, Editor and Chief, and author several chapters in the DLMF.

More than 35 years after the publication of A&S, Davis, then a professor at Brown, was invited back to hear about NIST's plans for the development of the DLMF. Davis' tepid response to our preliminary colorful, but static 3D graphs for the first DLMF chapter, Airy and related functions, actually sparked our research and design of 3D function surface visualizations that grew in sophistication as technology for displaying 3D graphics on the web advanced.

3.1. 3D Web Graphics. As mentioned earlier, we found that at the start of the DLMF project in the late 1990s and 2000s, the export of 3D graphics data by well-known computer algebra systems was inadequate for our needs. After looking into the graphics technology being used and studied at NIST, we concluded that VRML (Virtual Reality Modeling Language) was our best option. VRML is a 3D file format for creating interactive graphics for viewing on the web. There were a variety of VRML viewer browsers that could be freely downloaded, but over time the maintenance of some of the best was discontinued and the quality of new browsers was mixed. Furthermore, as we began to better understand what features we wanted to see in DLMF visualizations, the complexity of our visualizations increased, making it more and more difficult to find VRML browsers that could handle our graphics files, even when our codes appeared to follow VRML standards.

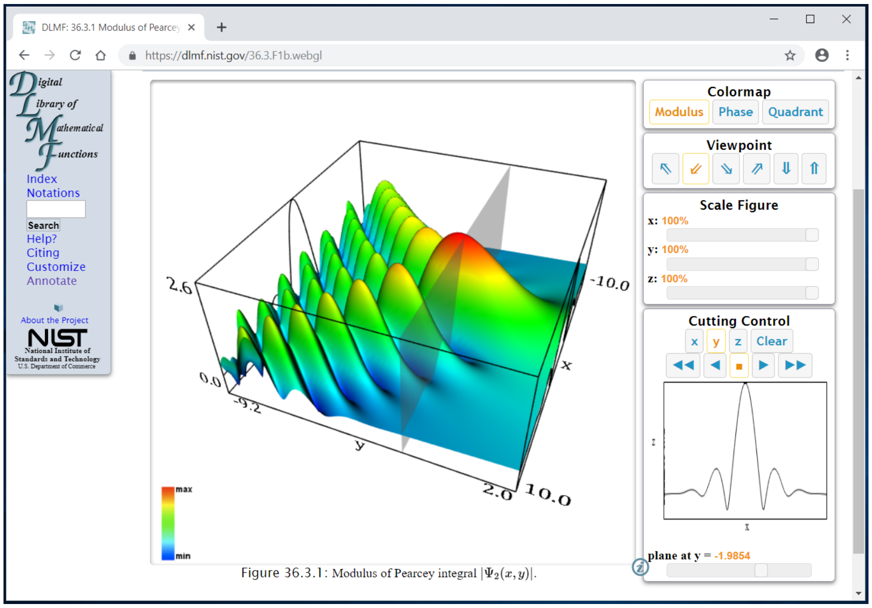
Noting the industry transition from VRML to X3D (Extensible 3D), graphics team member Qiming Wang designed a VRML to X3D converter. By the launch of the DLMF in 2010 we had created close to 200 interactive visualizations of mathematical function surfaces accessible in both formats [**12**].

However, our ultimate goal was to make the DLMF visualizations accessible on Windows, Mac, and Linux platforms. We found VRML/X3D browsers that worked for Windows and Mac, but never found a Linux browser that could successfully handle our graphics files. Also, having to download a viewer browser/plug-in to see the visualizations was a headache for both maintainers and users of the DLMF site. Problems arose whenever the browsers needed to be updated, or whenever there were changes to the platform operating system.

Motivated by these concerns, in mid 2011 we began monitoring the development of WebGL, a JavaScript API (application programming interface) for rendering graphics in a web browser without a viewer plug-in. Then, thanks to the work of Johannes Behr and colleagues [**2**] on X3DOM, which permitted the direct integration of X3D nodes into HTML content, we were able to make a crucial decision. We would convert all the DLMF visualizations to WebGL by using the X3DOM framework to build the application around our X3D codes. We were encouraged by our early success in creating a few initial visualizations that worked in a beta WebGL accessible Mozilla Firefox browser. We were also bolstered by preliminary X3DOM/WebGL work by NIST researcher Sandy Ressler and the work of Steven Birr, et al., on the LiverAnatomyExplorer WebGL Tool [**3**] . We began an intensive effort to convert all the DLMF 3D visualizations to WebGL and seamlessly integrate the displays into the HTML pages of the associated chapters. The new visualizations first appeared in DLMF Version 1.0.7 released on March 21, 2014.

Building our application using the X3DOM framework allowed us to reuse most of our X3D code to create the WebGL files. The most challenging work was recoding the dynamic displays and interactive features. After first creating stand alone WebGL files, we worked with NIST computer scientist Brian Antonishek and Bruce Miller, information architect of the DLMF website, to make the coding changes needed to integrate the visualizations into DLMF HTML files. We also made style changes to achieve a more polished look. Most importantly we successfully achieved our main goal: To reproduce or enhance the capabilities available in our VRML/X3D visualizations and provide additional capabilities where possible. WebGL is now the default format for viewing the DLMF visualizations and VRML/X3D files are being phased out [**19**].

3.2. DLMF Graphics Features. The best way to experience the DLMF visualizations is to go directly to the graphics sections found in most chapters and explore the capabilities available. The visualizations can now be viewed in most common web browsers on Windows, Mac, or Linux platforms. A few features are highlighted in this section.


DLMF: 36.3.1 Modulus of Pearcey
https://dlmf.nist.gov/36.3.F1b.webgl
Digital Library of Mathematical Functions
Index
Notations
Search
Help?
Citing
Customize
Annotate
About the Project
NIST
National Institute of Standards and Technology
U.S. Department of Commerce
2.6
0.0
-9.2
y
2.0
10.0
-10.0
x
max
min
Colormap
Modulus
Phase
Quadrant
Viewpoint
Scale Figure
Cutting Control
x
y
z
Clear
z
x

x: 100%

y: 100%

z: 100%

plane at y = -1.9854

Figure 36.3.1: Modulus of Pearcey integral $|\Psi_2(x,y)|$.

FIGURE 5. Modulus of Pearcey integral visualization embedded in DLMF webpage. Intersection of y direction cutting plane with surface displayed on bounding box sides and in pop-up display on side panel.

Figure 5 shows the general display for our surface visualizations. The user may click on the figure and rotate it freely or choose a stored viewpoint from the selection offered on the panel to the right. If a surface represents a complex valued function, the user is offered a phase, or argument, based color map in addition to the height, or modulus, color map. This option will not appear if the function is real valued.

Figure 6 shows a density plot for a Jacobian Elliptic function adjacent to its surface plot. A user can apply the scaling option to collapse the function in the vertical direction to obtain the density plot.

At the top of Figure 7 a type of Bessel function known as a Hankel function is shown with a height-based color map. Its branch cut is evident when one switches to a phase color map, and on scaling the surface height to zero, the phase density plot shown at the bottom emerges. On the website, one should note the difference in how the color spectrum is traversed as one travels around a pole versus a zero.

4. Current and Future Areas for Research

Clearly, a significant amount of work is involved in the design, creation, and maintenance of the visualizations in the DLMF. Initially, much of the 3D graphics work was motivated by deficiencies we saw in available software and computer algebra systems at the time. The rendering of 3D plots has improved in many

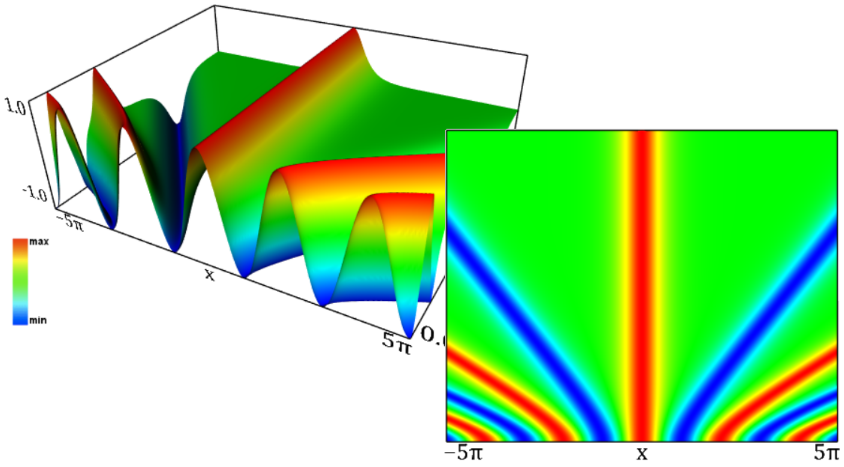


FIGURE 6. Jacobian Elliptic function $\mathrm{cn}(x, k)$ and density plot.

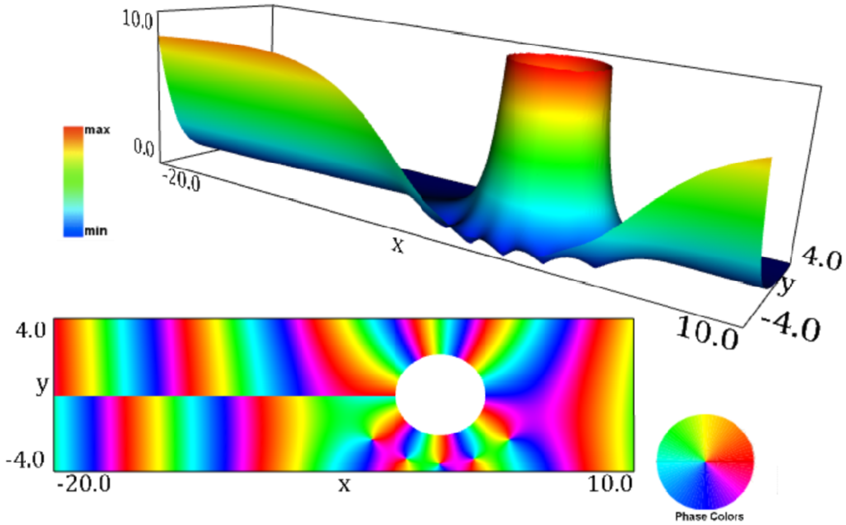


FIGURE 7. Modulus of Hankel function $H^{(1)}_{5.5}(x + iy)$ and phase density plot.

systems, and export options have expanded tremendously, but we still notice that the quality of the 3D data exported may not match what is seen on the screen.

Creating our own grids and visualizations gives us access to the data and routines that control our visualizations. This is helpful if we want to expand existing features or create new ones. Also, since our work is open to the public, we can get feedback from other researchers through publications and presentations at conferences.

There are several directions to go with our grid generation work. The ultimate goal is to develop a robust method that can be used to create quality grids in a reasonable amount of time. For now we will continue the work on adaptive curvature/gradient grids. We may also explore a parametric grid generation mapping which might work better if there are poles or other areas where there are steep gradients. Also, there have been some initial discussions with other grid generation researchers on the feasibility of creating a true zoom where the grid is refined and function values recomputed. Such an implementation would require a fast grid generation algorithm and hierarchical or locally refined techniques. Also, exploring unstructured triangulations and hybrid methods are still a possibility.

In addition to the true zoom, we might consider other changes to the visualizations such as adding or improving color maps, or including plots of real and imaginary parts of functions along with the modulus. In any case, while we expect to stick with our X3DOM/WebGL platform for the near future, we will strive to stay informed about trends in 3D web technology that might enhance our visualizations.

Disclaimer

All references to commercial products are provided only for clarification of the results presented. Their identification does not imply recommendation or endorsement by NIST.

References

1. M.A. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, NBS Applied Mathematics Series 55, National Bureau of Standards, Washington, DC, 1964.
2. J. Behr, P. Eschler, and M. Zollner, X3DOM: A DOM-based HTML5/X3D Integration Model, in *Proceedings of the 14th International Conference on 3D Web Technology(Web3D '09)*, ACM, S.N. Spencer, ed., pp. 127-135, 2009.
3. S. Birr, J. Monch,D. Sommerfeld, U. Preim, and B. Preim, The Liver Anatomy Explorer: A WebGL-based Surgical Teacing Tool, *IEEE Computer Graphics and Applications*, **33**, 5, pp.48-58, 2013.
4. C. de Boor, *A Practical Guide to Splines (revised edition)*, Springer, New York, 2001.
5. R.F. Boisvert, D.W. Lozier, Handbook of Mathematical Functions, in *A Century of Excellence in Measurements Standards and Technology*(D.R. Lide, ed.), CRC Press, pp. 135-139, 2001.
6. J. U. Brackbill, J. S. Saltzman, Adaptive zoning for singular problems in two dimensions, *J.Comput.Phys.* **46** (1982), 342-368.
7. T. D. DeRose, Composing Bezier simplexes, *ACM Transactions on Graphics* (3) **7** (1988), 198-221.
8. T. D. DeRose, R. N. Goldman, H. Hagen & S. Mann, Functional composition algorithms via blossoming, *ACM Transactions on Graphics* (1993), 113-135.
9. *NIST Digital Library of Mathematical Functions.* https://dlmf.nist.gov/, Release 1.022 of 2019-03-15. F.W.J. Olver, A.B. Olde Daalhuis, D.W. Lozier, B.I. Schneider, R.F. Boisvert, C.W. Clark, B.R. Miller, and B.V. Saunders, eds.
10. D. Gonsor, T. Grandine, A curve blending algorithm suitable for grid generation, in *Geometric Modeling and Computing: Seattle 2003*, Nashboro Press, Brentwood, 2004.
11. W.J. Gordon, C.A. Hall, Construction of Curvilinear Coordinate Systems and Applications to Mesh Generation, *International Journal for Numerical Methods in Engineering*, **7**, pp.461-477, 1973.

12. F.W.J. Olver, D.W. Lozier, R.F. Boisvert, C.W. Clark, A Special Functions Handbook for the Digital Age, *Notices Amer. Math Soc.* 58, 7, pp. 905-911, 2011.
13. B. V. Saunders, Algebraic grid generation using tensor product B-splines, *NASA CR-177968*, 1985.
14. B. V. Saunders, A boundary conforming grid generation system for interface tracking, *Computers Math. Applic.* **29** (1995), 1-17.
15. B. V. Saunders, Q. Wang, From 2d to 3d: numerical grid generation and the visualization of complex surfaces, in *Proceedings of the 7th International Conference on Numerical Grid Generation in Computational Field Simulations*, 51-60, Whistler, British Columbia, Canada, 2000.
16. B. V. Saunders, The application of numerical grid generation to problems in computational fluid dynamics, in *Council for African American Researchers in the Mathematical Sciences: Vol: III*, 95-106, Contemporary Mathematics Series **275**, American Mathematical Society, 2001.
17. B. V. Saunders, Q. Wang, From B-spline mesh generation to effective visualizations for the NIST digital library of mathematical functions, in *Curve and Surface Design: Avignon 2006*, 235-243, Nashboro Press, Brentwood, 2007.
18. B.V. Saunders, Q. Wang, B. Antonishek, Adaptive Composite B-Spline Grid Generation for Interactive 3D Visualizations, in *Proceedings of MASCOT12/ISGG2012(International IMACS Workshop and Bi-annual International Society for Grid Generation Conference)*, Las Palmas de Gran Canarias, 2012, IMACS Series in Computational and Applied Mathematics (Special Volume), 2014.
19. B. Saunders, B. Antonishek, Q. Wang, B. Miller, Dynamic 3d visualizations of complex function surfaces using X3DOM and WebGL, in *Proceedings of the 20th International Conference on 3D Web Technology (Web3D 2015)*, Crete, Greece, 219-225, ACM, New York, 2015.
20. B. Schneider, B. Miller, B. Saunders, NIST's Digital Library of Mathematical Functions, *Physics Today*, **71**(2):48–53, 2018, https://doi.org/10.1063/PT.3.3846.
21. B. K. Soni, Grid generation for internal flow configurations, *Computers Math. Applic.* **24** (1992), 191-201.
22. S. Steinberg, P. J. Roache, Variational grid generation, *Num. Meth. for P.D.E.s* **2** (1986), pp. 71-96, 1986.
23. J. F. Thompson, Z. U. A. Warsi & C. W. Mastin, *Numerical Grid Generation: Foundations and Applications*, North Holland, New York, 1985.

National Institute of Standards and Technology, 100 Bureau Drive, Stop 8910, Gaithersburg, MD 20899, USA
E-mail address: bonita.saunders@nist.gov

NIST SP 1283
August 2022

Proceedings of the ASME 2020
International Mechanical Engineering Congress and Exposition
IMECE2020
November 15-19, 2020, Portland, OR, USA

IMECE2020-23234

FORMALIZING PERFORMANCE EVALUATION OF MOBILE MANIPULATOR ROBOTS USING CTML

Omar Aboul-Enein[*]
Salisbury University
Salisbury, Maryland
National Institute of Standards
and Technology
Gaithersburg, Maryland
Email: omar.aboul-enein@nist.gov

Yaping Jing
Salisbury University
Salisbury, Maryland

Roger Bostelman
National Institute of Standards
and Technology
Gaithersburg, Maryland

ABSTRACT

Computation Tree Measurement Language (CTML) is a newly developed formal language that offers simultaneous model verification and performance evaluation measures. While the theory behind CTML has been established, the language has yet to be tested on a practical example. In this work, we wish to demonstrate the utility of CTML when applied to a real-world application based in manufacturing. Mobile manipulators may enable more flexible, dynamic workflows within industry. Therefore, an artifact-based performance measurement test method for mobile manipulator robots developed at the National Institute of Standards and Technology was selected for evaluation. Contributions of this work include the modeling of robot tasks implemented for the performance measurement test using Petri nets, as well as the formulation and execution of sample queries using CTML. To compare the numerical results, query support, ease of implementation, and empirical runtime of CTML to other temporal logics in such applications, the queries were re-formulated and evaluated using the PRISM Model Checker. Finally, a discussion is included that considers future extensions of this work, relative to other existing research, that could potentially enable the integration of CTML with Systems Modeling Language (SysML) and Product Life-cycle Management (PLM) software solutions.

1 INTRODUCTION

In the field of formal verification, Computation Tree Measurement Language (CTML) offers many advantages over conventional methods for conducting performance-reliability analysis. CTML combines performance and reliability evaluation capabilities under one language [1]. Jing and Miner establish the theoretical foundation for CTML, describe the advantages of the language, and provide an example of the language usage through modeling the Dining Philosophers problem. In their work, a few primary advantages of CTML are noted. First, CTML is able to respond to nested queries with either a real-valued quantity or a probability. In addition, CTML matches the functionality of Probabilistic Computation Tree Logic (PCTL) [2] and can respond to a (non-trivial) subset of Probabilistic Linear Temporal Logic (PLTL) queries that are not expressible in PCTL [1, 3]. Most importantly, the functionality of CTML extends beyond the aforementioned logics as the language can answer queries not expressible in either PCTL or PLTL [1]. For example, CTML can answer survivability queries, which ask how much time remains until an event occurs given that another event has already occurred. Finally, Jing and Miner theoretically established the running efficiency of CTML as polynomial with respect to the both the formula and state size [1].

The theoretical advantages of CTML just described also suggest potential benefit towards Industry 4.0 applications. Industry 4.0 is the convergence of "robotics, cyber-physical systems,

[*]Address all correspondence to this author.

Aboul-Enein, Omar; Jing, Yaping; Bostelman, Roger V. "Formalizing Performance Evaluation of Mobile Manipulator Robots using CTML." Presented at International Mechanical Engineering Congress and Exposition (IMECE 2020), Portland, OR, US. November 15, 2020 - November 19, 2020.

software services, and human participants" towards "interoperability, information transparency, technical assistance, and decentralized decisions" [4]. For example, consider the following work in assessing picker robot workflow (modeled using workflow nets) conformance to the specifications of an automated warehouse [4]. CTL queries were formulated to verify safety properties of the workflow, such as an absence of deadlocks, proper workflow completion, and proper workflow termination among other properties. CTML could theoretically extend this assessment to include how much time remains until a product is shipped given the robot just picked up the product or the probability a deadlock occurs within 15 minutes given that a procurement has just completed [1,4].

While the theoretical advantages of CTML have been tested through a dining philosopher example scenario and the original state-based language has been extended to support reasoning over paths with multiple actions and states, the adoption of CTML towards Industry 4.0 applications would benefit from a more focused exploration of its use in a real-world problem [1,5]. Therefore, in this work, we explore the viability of CTML and its theoretically established benefits by applying the language towards the formal verification of a contemporary, manufacturing-based application. The application consists of a performance measurement test method for mobile manipulators currently under development at the National Institute of Standards and Technology (NIST)[1] [6]. Mobile manipulation offers an abundance of potential real-world uses, especially within manufacturing environments. These applications include tasks such as sanding or painting large surfaces, welding large parts, managing conveyors, and other forms of flexible manufacturing [7]. It should also be noted that existing work has applied Model-Based Systems Engineering (MBSE) methodologies to the performance measurement test using Systems Modeling Language (SysML) [8]. Whereas the authors state the SysML model was verified through a systems review and referenced experiments, the utilization of CTML exemplifies simultaneous, probabilistic verification and performance evaluation of the test method procedure [8].

To formally evaluate the performance measurement test method, we selected the Petri-net formalism described in [9] to develop an initial, high-level model of the robot tasks performed as part of the mobile manipulator performance test method. We also selected and formulated a series of sample queries to be answered by CTML. We then adapted the model and queries for use with another existing tool, the Probabilistic Symbolic Model-Checker (PRISM) (Download at: `http://www.prismmodelchecker.org/download.php`) [10]. Following this step, we compared the supported queries, numerical results, Central Processing Unit (CPU) runtime, and ease of implementation. Finally, given the findings of the study, we discuss potential future applications of CTML, which includes integration with the previously mentioned SysML model of the performance test method and integration with Product Life-cycle Management (PLM) software.

2 BACKGROUND

2.1 About the Petri net Formalism

The Petri net formalism was selected to develop the initial model both for the high-level convenience of the formalism and for its ability to be procedurally converted to a format usable by CTML and PRISM. Petri nets are described as a "bipartite, directed graph populated by three types of objects" [9]. The three objects include places (graphically represented as hollow circles), transitions (represented by boxes or solid lines), and directed arcs. Places represent a possible state, condition, or available resource and transitions represent a change of state or action. Additionally, tokens (graphically represented as one or more filled circles or a variable on a place) represent the current fulfillment of a condition or availability of a resource. A "marking" consists of the distribution of tokens to places at any given time (with the "initial marking" at time zero). Arcs, which may be weighted, may only connect places to transitions or vice versa. Upon "firing" a transition, the number of tokens required by the weight of the input arc is consumed from the input place and the number of tokens indicated by the weight of the output arc is produced on the output place. A transition can only be fired, or is "enabled", if the input place contains enough tokens to satisfy the weight of the input arc. Inhibitor arcs (marked by a circle instead of an arrowhead) instead require that places have no tokens in order to fire [9].

2.2 The Mobile Manipulator Performance Test Method

The performance measurement test method modeled in this paper utilizes a Re-configurable Mobile Manipulator Artifact (RMMA) [6]. The RMMA provides a known, user-adjustable measurement uncertainty to compare to ground-truth. Additionally, The RMMA consists of an anodized, machined-aluminum build and has an adjustable height and table surface rotation. The table surface can be drilled and tapped with holes arranged in various geometric patterns (such as a square, circle, or sinusoid). The holes are analogous to points mounted with retro-reflective targets, which are digitally detected through the use of a retro-reflective laser sensor/emitter (RLS) mounted on the end-effector of a robotic arm.

The use of this non-contact registration sensor reduces the risk of collisions and is comparable to operations such as the classic peg-in-hole assembly task [11]. An example of an RMMA, RLS, and retro-reflective target is shown in Fig. 1a. 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.

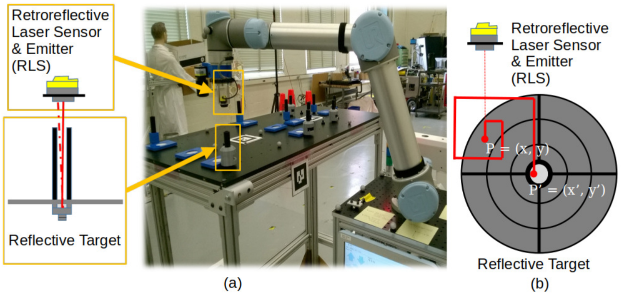

Retroreflective Laser Sensor & Emitter (RLS)
Reflective Target
(a)
Retroreflective Laser Sensor & Emitter (RLS)
P = (x, y)
P′ = (x′, y′)
Reflective Target
(b)

FIGURE 1. (a) THE RETRO-REFLECTIVE LASER SENSOR AND EMITTER (RLS) USED FOR ASSEMBLY POINT DETECTION. (b) TOP VIEW OF RLS AND SPIRAL SEARCH PATTERN

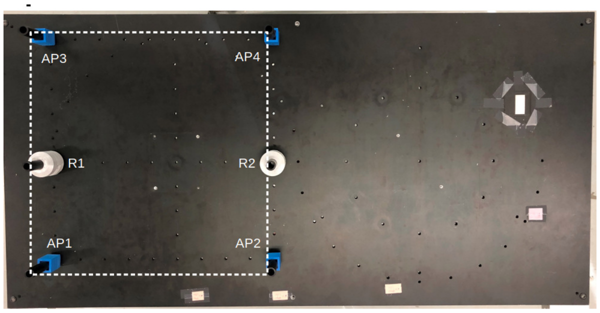


FIGURE 2. EXAMPLE OF RMMA CONFIGURATION WITH AP REFLECTORS IN A SQUARE PATTERN.

description of the general performance measurement test specifies two potential cases for manipulator and vehicle coordination. These levels of coordination are described as “indexed” (the manipulator performs assembly while vehicle is stopped) and “dynamic” (both the manipulator and the vehicle are in motion) [6]. The primary metrics of interest include a pass/fail test for initially detecting the retro-reflective targets at the assembly points, and a repeatability test in positioning the manipulator over the assembly points [6]. The RMMA can also evaluate and compare coordinate registration methods, and the time taken to locate the retro-reflective targets is also of interest for these methods [12].

3 PETRI NET MODEL DESCRIPTION

3.1 Model Scenario and Parameters

We now specify the instance of the test scenario to be modeled. First, we chose to focus on modeling a mobile manipulator utilizing an Automatic Guided Vehicle (AGV), specifically one that cannot re-compute paths around unexpected obstacles. With this system, an industrial manipulator arm is also fixtured onboard the AGV throughout the assembly tasks. Second is that we chose to model the indexed case of the mobile manipulator performance test instead of the dynamic case (see Section 2.2). Finally, we assume the RMMA is configured with six retro-reflective targets arranged in a square as shown in Fig. 2.

For the performance test method, the AGV travels to and parks at a sequence of *n* arbitrary locations next to the RMMA, which are referred to as “stops”. Each stop is assumed to be unique in either position and/or orientation. While in transit, it is possible that an obstacle or hazard causes the AGV to emergency stop (e-stop). Once the obstacle has been removed, the operator can clear the e-stop, allowing the AGV to resume the test. The AGV is allowed to e-stop *E* times before the entire test is aborted. While the AGV is parked at a stop, the manipulator arm first performs coordinate registration of the mobile manipulator to the RMMA by locating two retro-reflective targets, denoted R1 and R2. The manipulator arm then locates the other four retro-reflective targets, which are used as verification points to test the accuracy of the initial coordinate registration and manipulator positioning. These reflectors are denoted as assembly points 1 (AP1) through 4 (AP4). For both previous steps, the same target localization method is used, which consists of the manipulator tracing a square spiral pattern as visualized in Fig. 1b [12]. The spiral search is allowed to execute *t* steps in the localization pattern before the search is aborted. The detection of each assembly point, as part of the verification phase, can be repeated *r* times. Finally, this entire process can be repeated for *N* runs of the test. The values, or parameters, are adjustable settings of the performance test method. Changing each value as input to the model can increase or decrease the total number of reachable states associated with the Petri net.

3.2 Description of the Petri net Model

The model and initial marking of the mobile manipulator performance test (divided into five subnets) is shown in Fig. 3 - 7. Places are denoted P_i, where i is an integer subscript, and transitions are denoted t_j, where j is an integer subscript. Subnets, shown as rectangles, are annotated to indicate the connections between subnets. The exponential distribution was chosen as the firing distribution for transitions. The rate parameter of the exponential distribution was taken to be inversely proportional to the number of tokens currently in the Petri net at a given time.

3.2.1 AGV Subnet. The first subnet, shown in Fig. 3, defines the behavior of the AGV during the test. The initial token on P_1 represents the vehicle parked at the initial home position. After leaving the home position, the AGV proceeds through the sequence of *n* stops next to the RMMA. As mentioned previously, the AGV may e-stop while in transit (see Section 3.2.2). A token on P_2 indicates the AGV is finishing the current test run. P_3 maintains a count of the remaining number of test runs. P_3 initially has *N* tokens, which decreases every time t_2 is fired. For each run, the AGV parks at each of the n stops in sequence (see Section 3.2.3). When the token count on P_3 reaches 0, t_3 must fire, and the AGV attempts to return to the home position. A token on P_4 indicates the AGV has reached the home position and completed the test.

3.2.2 E-Stop Subnet. The subnet modeling the occurrence of an AGV e-stop is shown in Fig. 4. P_5 represents the vehicle en-route to a stop at the RMMA. P_6 is initialized with E tokens and counts the number of remaining e-stops allowed. P_7 counts the number of e-stops that have occurred previously. A token on P_8 indicates the vehicle is currently e-stopped, with an obstacle preventing the vehicle from currently reaching the intended destination. Since it is possible for the vehicle to recover from the e-stopped state, t_6 is included to connect P_5 to P_8.

3.2.3 Stop Subnet. The subnet modeling the actions of the AGV for each stop is shown in Fig. 5. A token on P_9 represents the AGV sending a signal to the manipulator to begin coordinate registration to the RMMA. P_{10} counts the number of times the AGV has previously visited the stop. A token on P_{11} indicates the AGV is currently parked at the RMMA. Transition t_8 passes the signal from the AGV to the manipulator and initializes r tokens on P_{18} in the manipulator subnet. t_9 passes a signal from the manipulator to the AGV indicating the manipulator has finished coordinate registration and verification. Upon receiving this token, the vehicle can proceed to the next stop.

3.2.4 Manipulator Subnet. The subnet defining the behavior of the manipulator arm is shown in Fig. 6. A token on P_{12} indicates the manipulator has received the signal from the AGV to begin coordinate registration. The manipulator begins in a stowed position, which allows for safe transport, and is represented by a token on P_{13}. Upon receiving the signal from the AGV, the manipulator moves to a stage position whereby it can access the RMMA. A token on P_{14} indicates that the manipulator is moving to the stage position. P_{15} and, similarly, P_{20} represent an e-stopped state. A token on P_{16} indicates the manipulator successfully reached the stage position. The manipulator then attempts the localization of two reflectors, R1 and R2, to complete coordinate registration. If either point cannot be found (as indicated by t_{26} in the spiral search subnet), the verification phase is skipped. If both points are localized, the manipulator proceeds to search for four verification reflectors, AP1 through AP4. A token on P_{17} indicates the end of a verification loop. P_{18} counts the remaining number of times that the verification loop will be repeated. A token on P_{19} represents the manipulator attempting to reach the stow position again. A token on P_{21} represents a signal being sent back to the AGV to proceed to the next stop.

3.2.5 Spiral Search Subnet. The subnet for modeling the spiral search localization method is pictured in Fig. 7, with Fig. 1b providing a visualization of the search routine itself. For the spiral search, a token on P_{22} indicates the manipulator in motion to position the RLS over an initial search point on the RMMA. The presence of a token on P_{23} represents a failure by the manipulator to position the toolpoint over the initial search point, which results in an e-stop. This is similarly the case for P_{30}. A token on P_{24} indicates the successful positioning of the RLS over the initial search point. Upon this success, t tokens are placed on P_{25}. P_{25} counts the remaining steps allowed in the search pattern. Upon reaching the initial search point, the RLS either immediately detects the retro-reflective target or does not. The detection of the target is indicated by a token on P_{26} while failing to detect the retro-reflective target is indicated by a token on P_{27}. If the target is not detected and search steps remain, the manipulator increments along the search path in attempt to localize the retro-reflective target. A token on P_{28} indicates the manipulator attempting to position the RLS over the next step along the search path. P_{29} counts the number of search steps previously taken by the manipulator.

4 EXPERIMENTS

4.1 Methods

We begin by explaining the intermediate steps taken to prepare the model and queries as input for the software implementation of CTML. Starting with the Petri net, we used the Stochastic Model-checking Analyzer for Reliability and Timing (SMART) (Download at: `https://asminer.github.io/smart/`) tool to convert the Petri net expression of the model to a Discrete Time Markov Chain (DTMC) [13]. Next, we used a custom-written Java program to express the DTMC in the format required by the CTML software. The Java program also added a set of "atomic functions" needed by the chosen queries [1]. The atomic functions calculate a starting quantity based on a given DTMC state. The augmented DTMC is then fed as input to the CTML software tool, along with the desired queries, to obtain the numerical result for each query (either a probability or a quantity). For this study, a total of ten models were generated from the Petri net by repeating this process with different input parameters. The parameters were varied to test a variety of model sizes ranging from a couple hundred states to approximately 4 million states.

For the two CTML queries that were expressible by PRISM, we were interested in comparing the numerical results, query support, ease of implementation, and CPU time between the two tools. To prepare the PRISM input, we started with the same DTMCs converted from the Petri net models via SMART. Unlike the CTML software, the PRISM tool required deadlocked states in the DTMC to have self-looping arcs, which increased the number of arcs for the PRISM input model (see Tab. 1). Since PRISM supports the specification of reward functions, which map states

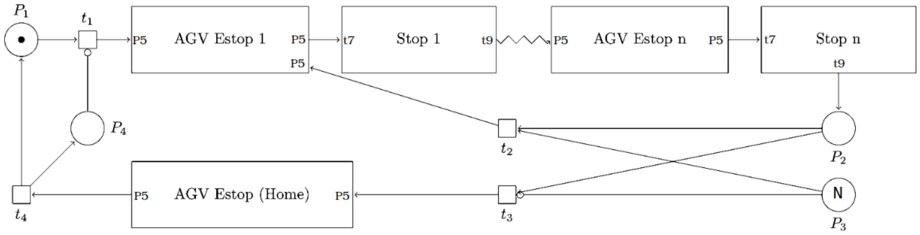


FIGURE 3. MOBILE MANIPULATOR PETRI NET: AGV SUBNET.

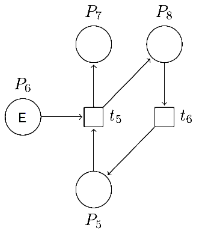


FIGURE 4. MOBILE MANIPULATOR PETRI NET: E-STOP SUBNET.

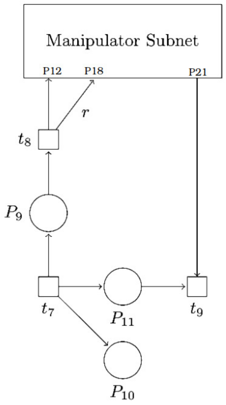


FIGURE 5. MOBILE MANIPULATOR PETRI NET: STOP SUBNET.

in the model to real values, each atomic function in CTML is redefined as a label (denoted by an "l" in its name) and a reward function, respectively [14]. Like CTML, the translated DTMC and queries can be fed into the PRISM tool as input.

The queries were tested on a computer with a 3.2GHz Intel Core i5 and 16GB of 1867MHz memory, and running MacOS X with the Java Development Kit (JDK) version 1.7. For CTML (a prototype tool) and PRISM (version 4.4), the runtime was benchmarked using the "User Time" field outputted by the `time` command line utility. This field is the total CPU time used by the tool excluding time used for executing operating system kernel code. The PRISM software tool also provided an internally measured time for model checking, which we called the "PRISM time" for short. To reduce the model-construction time in PRISM, we used "Explicit Model Files" and ran the queries using the explicit computation engine, which was specified using the `-ex` flag. With explicit model files, the states, transitions, and labels of the model were pre-constructed and each stored in a separate file. These models were then imported for model-checking by using the `-importmodel` flag. Additionally, to ensure there was enough memory for running the queries on the models, the maximum memory was expanded to 8 GB using the `-cuddmaxmem` and `-javamaxmem` flags [14]. All other PRISM configuration flags were kept to their default values.

4.2 Sample Queries

The sample queries were selected for their ability to test the boundaries of supported query types for existing temporal logics (as discussed in the Section 1) and for their relevance to the previously described metrics of the performance test method. Each query is presented in their English and CTML expressions. A complete explanation of CTML syntax and semantics is provided by Jing and Miner [1]. Likewise, for a complete explanation of the PRISM syntax and semantics, we refer to the PRISM manual [14]. Queries that could not be consistently translated into any of the logics supported by the PRISM tool are accompanied by an explanation. Otherwise, we present the query formulation in the appropriate logic. Finally, for all of the following queries, the atomic function *one* denotes a value of 1 for all states in the DTMC.

4.2.1 Query 1. *"What is the expected time until an e-stop occurs as the AGV attempts to park at stop 1?"* We define an atomic function $agvestop1$ with value 1 for the states where AGV is e-stopped, and 0 otherwise. Here, the atomic function *one* denotes one time unit per state. The query can be expressed

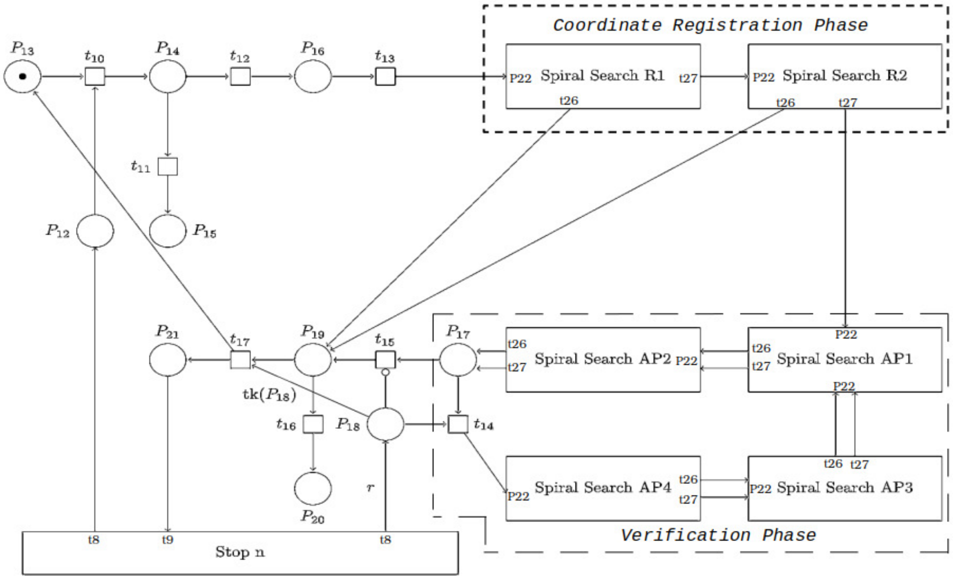


FIGURE 6. MOBILE MANIPULATOR PETRI NET: MANIPULATOR SUBNET.

in CTML as:

$$M\, one\, U_{+}\, agvestop1 \qquad (1)$$

The Until Plus operator U_{+} accumulates *one* along a path before reaching the first $agvestop1$ state, s, with $agvestop1(s) > 0$. The M operator denotes the expected value (mean) over the path formula $one\, U_{+}\, agvestop1$, which captures all paths reaching $agvestop1$ states. Attempting to evaluate Query 1 in PRISM resulted in several limitations and inconsistencies. First, attempting to import the reward functions while using the explicit engine resulted in an error stating there was no model generator to construct the rewards structure. This limitation was addressed as of PRISM version 4.6, however, due to the recent timing of this release, we were unable to directly test the newly supported functionality [15]. Therefore, without the explicit engine, the long model construction time discussed previously made it impractical to evaluate Query 1 on the larger models. With smaller models, for which the model construction time was not prohibitive, evaluation was attempted using the hybrid engine. In this case, PRISM returned a result of infinity, which was inconsistent with CTML. This behavior is a design choice in the PRISM property specification language and occurs when the probability of reaching a destination state is less than one [14, 16]. By specification of the mobile manipulator test scenario, it is not guaranteed that the AGV will e-stop. Therefore, Query 1 could not be fully evaluated using PRISM.

4.2.2 Query 2.

"What is the probability that the manipulator eventually fails to locate AP1 while the AGV is parked at stop 1?" For this query, we define an atomic function $ap1fail\text{-}stop1$ with value 1 for the states where the manipulator fails to locate AP1 while the AGV is parked at the first stop next to the RMMA, and 0 otherwise. This situation occurs if the maximum number of steps in the spiral search pattern has been exceeded and AP1 is still not detected by the RLS. This query can be expressed by CTML as:

$$M\, one\, U_{\times}\, ap1fail\text{-}stop1 \qquad (2a)$$

This is similar to Query 1, except that here we use the Until Muliply operator $U_{\times}$ to compute a probability along a path before reaching the first $ap1fail\text{-}stop1$ state s with $ap1fail\text{-}stop1(s) > 0$. The expected value operator M over the path formula $one\, U_{\times}\, ap1fail\text{-}stop1$ computes the sum over all states s with $ap1fail\text{-}stop1(s) > 0$, the probability that s is reached before any other state s' with $ap1fail\text{-}stop1(s') > 0$, multiplied by $ap1fail\text{-}stop1(s)$. Note that the CTML formula for Query 2 is shown in Eqn. 2a. Query 2 can be successfully translated to PCTL in PRISM using Eqn. 2b. The reachability operator F, as described in the PRISM manual, can be used to express formulas like "*true U* p", which specifies that p is eventually true. Then, the probability for the formula can be obtained via the operator $P_{=?}$ in PRISM.

$$P_{=?}\,(F\, ap1fail\text{-}stop1\_l) \qquad (2b)$$

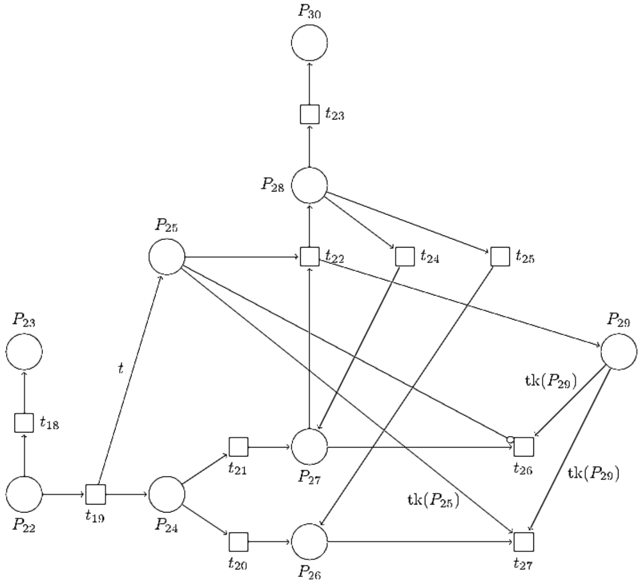


FIGURE 7. MOBILE MANIPULATOR PETRI NET: SPIRAL SEARCH SUBNET.

4.2.3 Query 3. *"Given that an AGV e-stop occurred while the AGV was attempting to reach stop 1, what is the probability that the manipulator fails to locate AP1?"* Similar to Query 6 presented by Jing and Miner for the dining philosopher model [1], Query 3 requires a conditional probability of the form $Pr(B|A)$, where A denotes an AGV e-stop while the AGV attempts to park at the first stop next to the RMMA. B denotes a failure to localize the AP1 reflector while the AGV is parked at the first stop next to the RMMA. To solve the conditional probability, we find the probability of events B and A occurring simultaneously and then divide the result by the probability that event A occurs. The CTML formula for this query is shown in Eqn. 3a, which can also be successfully translated to PLTL in PRISM and is shown in Eqn. 3b.

$$\frac{M\,one\,U_{\times}\,(\,(M\,one\,U_{\times}\,ap1fail\text{-}stop1\,) * agvestop1)}{M\,one\,U_{\times}\,agvestop1} \tag{3a}$$

$$\frac{P_{=?}\,(\,F\,(\,(agvestop1\_l\,) \wedge (F\,ap1fail\text{-}stop1\_l\,)\,)\,)}{P_{=?}\,(F\,agvestop1\_l\,)} \tag{3b}$$

4.2.4 Query 4. *"What is the expected number of steps taken to locate AP1 while the AGV is parked at stop 1?"* Here we define two atomic functions. $ap1step\text{-}stop1$ assigns a value of 1 for all states in which the manipulator increments an additional step in the spiral search localization pattern when attempting to locate the reflector AP1 while the AGV is parked at the first stop next to the RMMA, and 0 otherwise. The atomic function $ap1pass\text{-}stop1$ assigns a value of 1 for each state in which the manipulator detects the AP1 reflector while the AGV is parked at the first stop next to the RMMA, and 0 otherwise. The CTML formula for Query 4 is presented in Eqn. 4.

$$M\,ap1step\text{-}stop1\,U_{+}ap1pass\text{-}stop1 \tag{4}$$

For reasons similar to Query 1, and because the atomic function $ap1step\text{-}stop1$ does not have a match for the atomic proposition $true$ in PCTL, this query could not be translated for PRISM.

4.2.5 Query 5. *"Given the AGV is parked at stop 1 and the manipulator fails to locate AP1, what is the probability that AP2 is located within 10 time units?"* We define an atomic function $ap2pass\text{-}stop1$ that assigns a value of 1 for each state in which the manipulator detects the AP2 reflector while the AGV is parked at the first stop next to the RMMA, and 0 otherwise. This query is similar to Query 3, except we must use a time bounded formula. First, we determine the probability to reach the states where the AP2 reflector is located within 10 time steps when the AGV is parked at the first stop next to the RMMA, starting from every possible state. Then, we filter out all but the

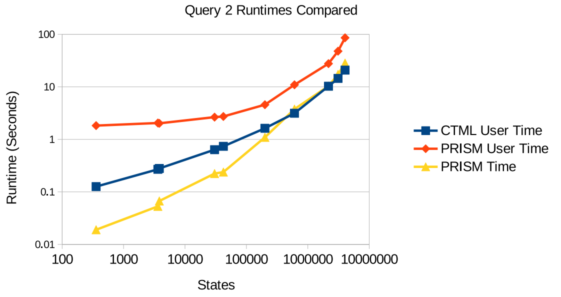


FIGURE 8. LINE PLOT COMPARING RUNTIMES OF QUERY 2.

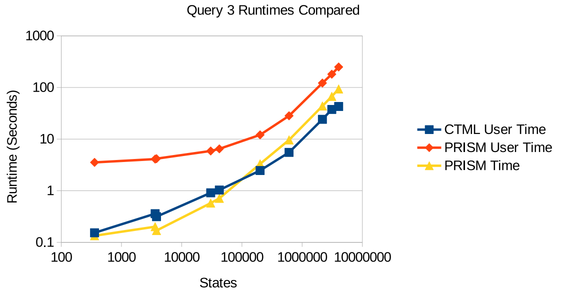


FIGURE 9. LINE PLOT COMPARING RUNTIMES OF QUERY 3.

states where the manipulator fails to locate AP1 while the AGV is parked at the first stop. We then sum, over all $ap1fail\text{-}stop1$ states, the probability to reach the first state multiplied by the probability to locate $ap2pass\text{-}stop1$ starting from the state. This quantity must be divided by $M\,one\,U_\times\ ap1fail\text{-}stop1$, which has been explained in Query 2. The CTML formula for Query 5 is shown in Eqn. 5.

$$\frac{M\,one\,U_\times((M\,one\,U_\times^{\leq 10}ap2pass\text{-}stop1) * ap1fail\text{-}stop1)}{M\,one\,U_\times\ ap1fail\text{-}stop1} \tag{5}$$

Query 5 could not be translated for use with the PRISM tool, since time-bounded operators are not expressible in PLTL [3,10].

4.2.6 Query 6.

"Given the AGV e-stops while the AGV is approaching stop 1 what is the expected time until the manipulator fails to locate AP1?" Similar to Queries 3 and 5, we first determine the expected time until the manipulator fails to locate the AP1 reflector when the AGV is parked at the first stop next to the RMMA, starting from each possible state. Then, we filter out all but the states where the AGV is e-stopped when attempting to reach the first stop next to the RMMA. We then sum over all $agvestop1$ states, the probability to reach the first state multiplied by the expected time to $ap1fail\text{-}stop1$ starting from that state. Finally, this quantity must be divided by $M\,one\,U_\times\ agvestop1$. The CTML formula for Query 6 is presented in Eqn. 6.

$$\frac{M\,one\,U_\times((M\,one\,U_+ap1fail\text{-}stop1) * agvestop1)}{M\,one\,U_\times\ agvestop1} \tag{6}$$

Query 6 could not be translated for use with PRISM because the tool does not support nested real-valued formulas [1].

5 RESULTS AND COMPARISON

The numerical results and CPU runtimes for each query evaluated using CTML and PRISM are listed in Tab. 1, which shows the relative precision of 10^{-4} for numerical results. Line plots comparing the runtimes between CTML and PRISM for Queries 2 and 3 are also provided in Fig. 8 and Fig. 9.

We now interpret the CTML results of the sample queries and compare the difference in query support between the two tools. The results for Query 1 yielded one time-step on average until the AGV e-stops as the AGV travels to the first stop next to the RMMA. When we compare the results of Queries 2 and 3, the probability of failing to detect AP1 is not influenced by the occurrence of an AGV e-stop since the result of both queries tend toward zero. It is also shown that, in absence of additional uncertainty, failing to detect the AP1 reflector is not a likely occurrence. This is further reinforced by the results of Query 4, which shows that between no steps and one step on average should be required to detect AP1 when the AGV is parked at the first stop. Query 5, however, shows that the probability of failing to detect AP1 at the first stop and then detecting AP2 within 10 times steps is 37%. The result for Query 6 shows that, given an AGV e-stop occurred, between 0 and 1 time-steps elapse on average until the manipulator fails to locate AP1. Out of the six queries evaluated using CTML, only two could be translated such that all of the generated models could be successfully and consistently evaluated using PRISM for the reasons given in Section 4.2. For Queries 2 and 3, the numerical results of PRISM were identical to those of CTML.

For the ease of implementation, preparing the model as input for PRISM was more complicated than CTML since the states, transitions, and labels of the model had to be split across three files for each generated DTMC model whereas CTML could define all needed structures in one file. Furthermore, CTML did not require separately defined reward and labeling functions, as was the case with PRISM. We were also unable to determine a way to explicitly define different, named reward functions for use with the property specification language. This restricted us to using only one explicitly-defined reward function at a time.

In terms of the runtime, the longest user time for executing a query in CTML corresponded with Query 1, which was evaluated

TABLE 1. SAMPLE QUERY RESULTS AND CPU RUN TIME

		Parameters ($E\ t\ r\ n\ N$)[2]									
		1, 1, 1, 1, 1	2, 2, 1, 2, 1,	4, 10, 1, 1, 1	4, 10, 1, 2, 1	3, 3, 1, 3, 1	4, 10, 1, 3, 1	4, 4, 1, 4, 1	4, 10, 1, 4, 2	4, 10, 1, 4, 3	4, 10, 1, 4, 4
		States									
		357	3,621	3,819	30,305	42,164	200,115	606,305	2,171,190	3,106,315	4,041,440
		Arcs (CTML)									
		421	4,352	4,702	37,384	51,139	247,054	739,504	2,683,829	3,840,954	4,998,079
		Arcs (PRISM)									
Query No.	**Tool**	501	5,189	5,637	44,789	61,055	295,909	883,669	3,212,934	4,597,559	5,982,184
Q1[3]	CTML	1.1118	1.0008	1.0011	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
User Time[4]		0.140	0.361	0.358	0.980	0.995	4.116	7.069	57.041	77.943	127.787
Q2a[5]	CTML	0.0049	0.0018	0.0000	0.0000	0.0006	0.0000	0.0002	0.0000	0.0000	0.0000
User Time		0.126	0.269	0.279	0.636	0.737	1.624	3.144	10.276	14.478	20.812
Q2b[5]	PRISM	0.0049	0.0018	0.0000	0.0000	0.0006	0.0000	0.0002	0.0000	0.0000	0.0000
User Time		1.829	2.050	2.020	2.652	2.729	4.568	10.935	27.677	47.722	85.786
PRISM Time[6]		0.019	0.053	0.067	0.223	0.238	1.094	3.761	10.350	17.231	28.285
Q3a[5]	CTML	0.0049	0.0018	0.0000	0.0000	0.0006	0.0000	0.0002	0.0000	0.0000	0.0000
User Time		0.153	0.359	0.314	0.911	1.031	2.461	5.521	24.174	37.526	42.835
Q3b[5]	PRISM	0.0049	0.0018	0.0000	0.0000	0.0006	0.0000	0.0002	0.0000	0.0000	0.0000
User Time		3.536	4.097	4.176	5.888	6.472	12.085	28.287	122.074	180.884	249.742
PRISM Time		0.134	0.201	0.169	0.575	0.711	3.3000	9.6280	43.7320	67.1710	93.4110
Q4[7]	CTML	0.0049	0.0091	0.0132	0.0132	0.0115	0.0132	0.0125	0.0132	0.0132	0.0132
User Time		0.143	0.321	0.298	0.758	1.012	2.471	6.122	27.254	34.247	45.738
Q5[5]	CTML	0.3333	0.3611	0.3704	0.3704	0.3704	0.3704	0.3704	0.3704	0.3704	0.3704
User Time		0.153	0.313	0.305	0.908	1.176	2.494	4.675	18.340	23.682	34.548
Q6[3]	CTML	0.0865	0.0386	0.0000	0.0000	0.0153	0.0000	0.0057	0.0000	0.0000	0.0000
User Time		0.163	0.370	0.387	1.011	1.097	4.730	8.783	41.010	55.517	68.300

[2] See Section 3.1; [3] Computed as time-steps; [4] User Mode CPU Time, in seconds, as measured by the `time` command line utility; [5] Computed as probability; [6] Internally measured model-checking time, in seconds, offered by PRISM; [7] Computed as spiral search-steps.

in just over 2 minutes when run on a model with approximately 4 million states and almost 5 million arcs. When comparing the runtimes of Queries 2 and 3, it was found that the user time of CTML was consistently lower than the user time measured for PRISM as the number of states and arcs in the models increased. However, the internally measured PRISM model checking time, or "PRISM time", yielded the following. For Query 2, the PRISM time was slightly lower than the CTML user time for the models with less than 600,000 states. For models exceeding 600,000 states, the PRISM time exceeded the CTML user time. For Query 3, the CTML user time overtook the PRISM time after the model size exceeded 200,000 states.

6 CONCLUSION

In this work, we explored the practical utility of the theoretically established advantages of CTML by using the language to evaluate a set of sample queries on a modeled, manufacturing-based application. The benefits and practical application of CTML established in this paper allows for consideration of future extensions. For example, existing work has pursued the development of a formal verification framework for SysML activity diagrams in which activity diagrams were translated to Probabilistic Timed Automata (PTAs) for use with PRISM [17]. Alternative approaches opted to map the activity diagrams to DTMCs or Petri nets [18,19]. Integration of CTML into such frameworks could further enhance CTML ease of implementation by allowing the use of existing SysML models such as the previously developed mobile manipulator performance test model [8].

Furthermore, such work could facilitate an exploration towards integrating CTML with PLM software. Existing work has demonstrated interest in unifying MBSE with PLM with noted benefits including the ability of such models to adapt throughout the design process and to provide a centralized source of design reference [20]. Interest has also been shown in efforts to unify the use of temporal-logic based verification and MBSE within PLM tool chains [21,22]. In this work, Form-L language (described as being comparable to LTL) was integrated with the Modelica language to facilitate subsequent case study on requirements management for cyber-physical systems [21].

REFERENCES

[1] Jing, Y., and Miner, A. S., 2018. "Computation tree measurement language (CTML)". *Formal Aspects of Computing,* ***30***(3-4), August, pp. 443–462.

[2] Hansson, H., and Jonsson, B., 1994. "A logic for reasoning about time and reliability". *Formal Aspects of Computing,* ***6***(5), September, pp. 512–535.

[3] Costas, C., and Mihalis, Y., 1995. "The complexity of probabilistic verification". *Journal of the ACM,* ***42***(4), July, pp. 857–907.

[4] Kattepur, A., 2019. "Workflow composition and analysis in industry 4.0 warehouse automation". *IET Collaborative Intelligent Manufacturing,* ***1***(3), October, pp. 79–89.

[5] Jing, Y., and Miner, A. S., 2018. "Action and state based computation tree measurement language and algorithms". In 15th International Conference on Quantitative Evaluation of Systems, A. McIver and A. Horvath, eds., Vol. 11024 of *Lecture Notes in Computer Science*, Springer Cham, pp. 190–206.

[6] Bostelman, R. V., Hong, T., and Marvel, J. A., 2015. "Performance measurement of mobile manipulators". In Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications, J. J. Braun, ed., Vol. 9498, International Society for Optics and Photonics, SPIE, pp. 97 – 106.

[7] Bostelman, R. V., Hong, T., and Marvel, J., 2016. "Survey of research for performance measurement of mobile manipulators". *Journal of Research (NIST JRES),* ***121***, June, pp. 342–366.

[8] Bostelman, R. V., Foufou, S., Hong, T., and Shah, M., 2017. "Model of mobile manipulator performance measurement using sysml". *Journal of Intelligent and Robotic Systems,* ***92***, September, pp. 65–83.

[9] Zurawski, R., and Zhou, M., 1994. "Petri nets and industrial applications: A tutorial". *IEEE: Transactions on Industrial Electronics,* ***41***(6), December, pp. 567–583.

[10] Kwiatkowska, M., Norman, G., and Parker, D., 2011. "PRISM 4.0: Verification of probabilistic real-time systems". In 23rd International Conference on Computer Aided Verification, G. Gopalakrishnan and S. Qadeer, eds., Vol. 6806 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 585–591.

[11] Bostelman, R. V., Foufou, S., Legowik, S. A., and Hong, T. H., 2016. "Mobile manipulator performance measurement towards manufacturing assembly tasks". In Product Lifecycle Management for Digital Transformation of Industries - 13th IFIP WG 5.1 International Conference, PLM 2016, Revised Selected Papers, B. Eynard, A. Bouras, A. Bernard, L. Rivest, and R. Harik, eds., Vol. 492 of *IFIP Advances in Information and Communication Technology*, Springer New York LLC, pp. 411–420.

[12] Bostelman, R. V., Eastman, R., Hong, T. H., Aboul-Enein, O., Legowik, S. A., and Foufou, S., 2016. "Comparison of registration methods for mobile manipulators". In Advances in Cooperative Robots, pp. 205–213.

[13] Ciardo, G., and Miner, A. S., 2004. "Smart: the stochastic model checking analyzer for reliability and timing". In 1st International Conference on Quantitative Evaluation of Systems, IEEE, pp. 338–339.

[14] Parker, D., Norman, G., and Kwiatkowska, M., 2019. Prism manual. On the WWW, April. URL `http://www.prismmodelchecker.org/manual/Main/AllOnOnePage`.

[15] Parker, D., Norman, G., and Kwiatkowska, M., 2020. Prism changelog. On the WWW, April. URL `https://github.com/prismmodelchecker/prism/blob/master/CHANGELOG.txt`.

[16] Kwiatkowska, M., 2007. "Quantitative verification: models techniques and tools". In Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering, ESEC-FSE '07, Association for Computing Machinery, pp. 449–458.

[17] Baouya, A., Bennouar, D., Mohamed, O. A., and Ouchani, S., 2015. "A quantitative verification framework of sysml activity diagrams under time constraints". *Expert Systems with Applications,* ***42***(21), November, pp. 7493–7510.

[18] Jarraya, Y., Soeanu, A., and Debbabi, M., 2007. "Automatic verification and performance analysis of time-constrained sysml activity diagrams". In 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems, IEEE, pp. 515–522.

[19] Huang, E., McGinnis, L. F., and Micthell, S. W., 2019. "Verifying sysml activity diagrams using formal transformation to petri nets". *Systems Engineering,* ***23***(1), November, pp. 118–135.

[20] Bajaj, M., Zwemer, D., Yntema, R., Phung, A., Kumar, A., Dwivedi, A., and Waikar, M., 2016. "Mbse++ — foundations for extended model-based systems engineering across system lifecycle". *INCOSE International Symposium,* ***26***(1), September, pp. 2429–2445.

[21] Garro, A., Tundis, A., Bouskela, D., Jardin, A., Thuy, N., Otter, M., Buffoni, L., Fritzson, P., Sjölund, M., Schamai, W., and Olsson, H., 2016. "On formal cyber physical system properties modeling: A new temporal logic language and a modelica-based solution". In 2016 IEEE International Symposium on Systems Engineering, IEEE, pp. 1–8.

[22] Aiello, F., Garro, A., Lemmens, Y., and Dutré, S., 2017. "Simulation-based verification of system requirements: An integrated solution". In 2017 IEEE 14th International Conference on Networking, Sensing and Control, pp. 726–731.

2020 IEEE/ACM Symposium on Edge Computing (SEC)

Task Management for Cooperative Mobile Edge Computing

Li-Tse Hsieh, Hang Liu
The Catholic University of America
Washington, DC, USA

Yang Guo
National Institute of Standards and Technology
Gaithersburg, MD, USA

Robert Gazda
InterDigital Communications, Inc.
Conshohocken, PA, USA

Abstract—**This paper investigates the task management for cooperative mobile edge computing (MEC), where a set of geographically distributed heterogeneous edge nodes not only cooperate with remote cloud data centers but also help each other to jointly process tasks and support real-time IoT applications at the edge of the network. Especially, we address the challenges in optimizing assignment of the tasks to the nodes under dynamic network environments when the task arrivals, node computing capabilities, and network states are non-stationary and unknown a priori. We propose a novel stochastic framework to model the interactions of the involved entities, including the edge-to-edge horizontal cooperation and the edge-to-cloud vertical cooperation. The task assignment problem is formulated and the algorithm is developed based on online reinforcement learning to optimize the performance for task processing while capturing various dynamics and heterogeneities of node computing capabilities and network conditions with no requirement for prior knowledge of them. Further, by leveraging the structure of the underlying problem, a post-decision state is introduced and a function decomposition technique is proposed, which are incorporated with reinforcement learning to reduce the search space and computation complexity. The evaluation results demonstrate that the proposed online learning-based scheme outperforms the state-of-the-art benchmark algorithms.**

Keywords-mobile edge computing (MEC); task assignment; stochastic optimization; reinforcement learning; decomposition

I. INTRODUCTION

The convergence of communication technologies, information processing, embedded systems, and automation has enabled rapid growth of the Internet of Things (IoT). Various things or objects such as sensors, actuators, and smart devices are connected to the Internet to provide new services such as smart cities, intelligent transportation, and industrial control. These emerging applications often involve performing intensive computations on sensor data, e.g. image/video in real time, aiming to realize fast interactions with the surrounding physical world. Mobile edge computing (MEC) has been advocated to support real-time IoT applications. Edge nodes with computing, storage and communication capabilities are co-located or integrated with base stations (BSs), routers, and gateways in the mobile radio access network (RAN) to execute sensor data processing tasks, such as image recognition and object detection, near the data sources at the edge of the network. Compared to the traditional cloud-based solutions, MEC can reduce data transfer time and conserve communication bandwidth by not shipping large volumes of data collected from many sensors to a centralized data center over the Internet, while providing real-time local context-aware services required by emerging IoT applications.

In contrast to centralized cloud data centers, MEC edge nodes are deployed at geographically distributed locations in a RAN, and user requests for computational tasks may arrive at any MEC edge node, instead of a gateway or master node. Individually, edge nodes have limited and heterogeneous computing resources as well as dynamic network conditions. The tasks may be queued at an edge node due to its limited processing capability and even dropped due to the node's bounded buffer. In addition, the workload received by edge nodes exhibits temporal and spatial fluctuations due to the bursty nature of IoT applications and mobility. If edge nodes can forward the unprocessed tasks to nearby edge nodes and/or remote cloud data centers for execution, the overall processing capability will be increased. The horizontal cooperation among edge nodes as well as the vertical cooperation between edge nodes and remote cloud for jointly processing computational tasks can balance the workload and reduce service latency. However, there are non-trivial challenges to manage the MEC services and assign the tasks to be processed at different nodes in a distributed and dynamic MEC network to achieve the optimal system performance: a) both computing resource availability at a node and network communication delay between the nodes should be taken into consideration to make the best task assignment decision for forwarding tasks from one node to another. b) The task arrivals, available computing capabilities at edge nodes, and network delays are time-varying and unknown a priori in many MEC scenarios.

Most research efforts have focused on the problem of offloading tasks from mobile devices to edge nodes [1], [2] or the vertical cooperation in which MEC edge nodes help cloud data centers process delay-sensitive tasks for improved quality of service (QoS) [3], [4]. There are less attentions to investigate the horizontal cooperation among MEC edge nodes for joint task processing. Recently, the authors in [5] proposed an offloading scheme that allows an edge node to forward its tasks to other edge nodes for processing to balance the workload. However, they assume that users submit their tasks to edge nodes at a constant rate and the task arrival rate at an edge node is known. The queuing delay at an edge node and the network delay between the edge nodes are also deterministic and can be known in advance. These

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Task Management for Cooperative Mobile Edge Computing." Presented at Fifth ACM/IEEE Symposium on Edge Computing: Third Workshop on Hot Topics on Web of Things. November 11, 2020 - November 13, 2020.

assumptions are too idealized for real deployment scenarios. Furthermore, their task assignment algorithm is based on classical convex optimization methods given a static MEC environment, which fails to characterize system dynamics and impacts the performance.

In this paper, we investigate the task assignment and management for cooperative mobile edge computing services under time-varying task arrivals, node computing capabilities, and network states. We cast the task assignment as a dynamic and stochastic optimization problem and develop an online reinforcement learning algorithm to fully explore the synergy among the MEC entities and achieve optimal QoS performance with no assumption on prior knowledge of the underlying network dynamics. Specifically, we propose a novel stochastic framework to model the horizontal cooperation of edge nodes as well as the vertical cooperation between edge nodes and cloud data centers, and capture various dynamics and heterogeneity of node computation capabilities and MEC network conditions. The task assignment problem is formulated as a Markov decision process (MDP). The optimization algorithm is then developed based on online reinforcement learning. In order to reduce the computational complexity and to improve the learning algorithm efficiency, we propose post-decision state estimation and function decomposition techniques by leveraging structure of the underlying problem. Numerical results show that our proposed approach improves the MEC network performance, compared to the existing algorithms. To the best knowledge of the authors, this is the first work to solve the task assignment optimization problem with edge-to-edge horizontal cooperation and edge-to-cloud vertical cooperation under stochastic and dynamic MEC network environments by employing a machine learning-based approach.

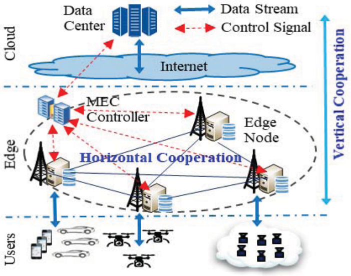


Figure 1. System model.

II. System Model and Problem Formulation

A. System Model

In this paper, we consider a software-defined MEC network with a centralized control plane and a distributed data plane [6]. Software-defined networks (SDNs) have attracted a lot of interest from network service providers because they can be flexibly controlled and programmed. As shown in Fig. 1, a MEC network consists of geographically distributed edge nodes deployed in a RAN covering a certain area. The edge nodes are equipped with computing resources and co-located or integrated with base stations or WiFi access points. They connect to a cloud data center through the Internet. We consider the data center as a special node with powerful resources but far from the RAN. Smart devices/sensors connect to nearby MEC edge nodes to submit their computational tasks, e.g. analyzing sensed video data. The MEC nodes (edge nodes and data center) help each other to jointly process the computational tasks. When an edge node receives the tasks from its associated smart devices, it either process them locally, or forward part or all of its unprocessed tasks to other edge nodes or to the cloud data center for processing to optimize the QoS, which is based on the task assignment decision. In the SDN-based MEC network, a control plane connects the edge nodes to a software-defined programmable MEC controller that makes the task assignment decisions by taking into consideration the network and workload conditions. The MEC controller resides in the RAN and could be one of the edge nodes with dedicated control plane connectivity, thus the control latency is minimal.

Consider a MEC network that consists of *N* edge nodes, labeled as $\mathcal{N}$ = {1, 2, …, *N*} and a remote cloud data center modeled as a special node n_c. Note that it can be easily extended to multiple data centers. We assume that the system operates over discrete scheduling slots of equal time duration. The values of a two-dimensional task assignment matrix $\boldsymbol{\Phi}^t = \{\phi^t_{n,j}: n, j \in \mathcal{N} \cup n_c\}$ are decided at the beginning of each time slot *t*, where $\phi^t_{n,j}$ specifies the number of tasks that edge node *n* will send to edge node *j* or cloud data center n_c for processing in slot *t*, and $\phi^t_{n,n}$ is the number of tasks that edge node *n* will buffer for processing by itself. $\boldsymbol{\phi}^t_n = \{\phi^t_{n,j}, \phi^t_{j,n} : j \in \mathcal{N} \cup n_c\}$ represents the task assignment vector regarding edge node *n*. We assume that the data center n_c will process all the received tasks by itself, not forwarding them to the edge nodes, i.e. $\phi^t_{n_c,j} = 0, j \in \mathcal{N}$.

B. Problem Formulation

We first formulate the problem of stochastic task assignment optimization and then discuss the approaches to solve the optimization problem. Let A^t_n be the number of the new tasks randomly arrived at edge node *n*, $n \in \mathcal{N}$ from its associated devices in time slot *t*, and $\boldsymbol{A^t}$= $\{A^t_n : n \in \mathcal{N}\}$. The distribution of A^t_n is not known beforehand. Q^t_n represents the task queue length of node *n* at the beginning of time slot *t*. Let s^t_n be the task processing capability of node *n* in slot *t*, which is defined as the maximal number of tasks that node *n* can serve in slot *t*. We assume that s^t_n varies in time and is also unknown a priori. The queue evolution of node *n* can be written as,

$$Q^{t+1}_n = \begin{cases} 0, & if\ s^t_n \geq Q^t_n + \mathrm{A}^t_n + \Sigma_{i\in e_n}\phi^t_{i,n} - \Sigma_{i\in e_n}\phi^t_{n,i} \\ \min\{Q^t_n + \mathrm{A}^t_n + \Sigma_{i\in e_n}\phi^t_{i,n} - \Sigma_{i\in e_n}\phi^t_{n,i} - s^t_{n,}, Q^{(max)}_n\}, & \text{otherwise} \end{cases} \tag{1}$$

where $Q^{(max)}_n$ is the maximum queue buffer size at node *n*. An edge node may forward some of its tasks to other nodes for processing, or offer help to process the tasks from other nodes. $\Sigma_{i\in e_n}\phi^t_{n,i}$ where $e_n = \{\mathcal{N} \cup n_c\}\backslash\{n\}$ is the number of tasks that edge node *n* offloads to other nodes, and $\Sigma_{i\in e_n}\phi^t_{i,n}$ is the

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Task Management for Cooperative Mobile Edge Computing." Presented at Fifth ACM/IEEE Symposium on Edge Computing: Third Workshop on Hot Topics on Web of Things. November 11, 2020 - November 13, 2020.

number of tasks that edge node *n* receives from other nodes in slot *t*.

The local state of a node is characterized by its task queue size, its task processing capability, and its network delay to other nodes. For a node $n, n \in \mathcal{N} \cup n_c$, at the beginning of time slot *t*, we measure its local state as $\boldsymbol{\chi}_n^t = (Q_n^t, s_{n,}^t \boldsymbol{c}_n^t)$ where $\boldsymbol{c}_n^t = \{c_{n,j}^t, c_{j,n}^t : j \in \mathcal{N} \cup n_c\}$ with $c_{n,j}^t$ being the network delay for shipping a task from node *n* to node *j*, $c_{j,n}^t$ being to the network delay for shipping a task from node *j* to node *n*, and $c_{n,n}^t = 0$. As the network delay between two nodes is related to the transmission distance (the number of hops along the path between the two nodes), traffic conditions in the network, and many other unpredicted factors, it varies in time and its distribution is unknown a priori as well. At the beginning of each scheduling time slot *t*, the global MEC network state is represented $\boldsymbol{\chi}^t = \{\boldsymbol{\chi}_n^t : n \in \mathcal{N} \cup n_c\} = (\boldsymbol{Q}^t, \boldsymbol{S}^t, \boldsymbol{C}^t) \in X$, where $\boldsymbol{Q}^t = \{Q_n^t : n \in \mathcal{N} \cup n_c\}$, $\boldsymbol{S}^t = \{s_n^t : n \in \mathcal{N} \cup n_c\}$, and $\boldsymbol{C}^t = \{\boldsymbol{c}_n^t : n \in \mathcal{N} \cup n_c\}$. X represents the whole MEC system state space.

For a given MEC network state $\boldsymbol{\chi}^t$ at the beginning of a time slot *t*, a task assignment $\boldsymbol{\Phi}^t = \boldsymbol{\Phi}(\chi^t) = [\phi_{n,j}(\boldsymbol{\chi}^t): n, j \in \mathcal{N} \cup n_c\}$ is made, and the MEC network achieves an instantaneous utility that is related to the QoS. We consider delay-sensitive applications, where the QoS is measured by the task service delay and the task drop rate. The task service delay is defined as the period from the time that a task arrives at an edge node to the time that the task has been served in the unit of scheduling slot duration. For an edge node n, $n \in \mathcal{N}$, its service delay d_n depends on the delay incurred by the queue Q_n if edge node *n* processes the task by itself, or consists of the network delay $c_{n,j}$ and the queueing delay due to the queue Q_j at the service provider *j* if a task is sent from node *n* to node *j* for processing. The task drop rate o_n is defined as the number of tasks dropped per time slot due to buffer overflow.

The instantaneous MEC network utility under the state $\boldsymbol{\chi}^t$ and task assignment decision $\boldsymbol{\Phi}(\chi^t)$ at time slot *t* is defined as,

$$U(\chi^t, \boldsymbol{\Phi}(\chi^t)) = \sum_{n \in \mathcal{N}} [w_d U_n^{(d)}(\boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t)) + w_o U_n^{(o)}(\boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t))] \tag{2}$$

where $U_n^{(d)}(.)$ and $U_n^{(o)}(.)$ measure the satisfactions 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 utility function of the MEC system, respectively. For an edge node, we consider there is a maximal tolerance threshold, $d^{(\max)}$ for the service delay, i.e. $d_n \leq d^{(\max)}$. Correspondingly, let $o^{(\max)}$ be the maximal tolerance threshold for the task drop rate, i.e. $o_n \leq o^{(\max)}$. In addition, we choose the utility function to be the exponential functions, namely $U_n^{(d)} = \exp(-d_n/d^{(\max)})$ and $U_n^{(o)} = \exp(-o_n/o^{(\max)})$ [7].

Stochastic task arrivals and dynamic MEC system states present challenges and make traditional one-shot deterministic optimization schemes unstable and unable to achieve the optimal network performance on a longer timescale. Therefore, we want to develop a stochastic optimization framework for the cooperative task assignment, which maximizes the expected long-term utility of a MEC system while guaranteeing the service delay and task drop rate are within their respective acceptable thresholds.

The task assignment matrix $\boldsymbol{\Phi}(\boldsymbol{\chi}^t)$ is determined according to the control policy $\boldsymbol{\Phi}$ after observing the network state $\boldsymbol{\chi}^t$ at the beginning of a time slot *t*. The task assignment policy $\boldsymbol{\Phi}$ then induces a probability distribution over the set of possible MEC network states $\boldsymbol{\chi}^{t+1}$ in the following time slot, and hence a probability distribution over the set of per-slot utility $U(\boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t))$. For simplicity, we assume that the probability of a state in the subsequent slot depends only on the state attained in the present slot, i.e. the task processing capability of a node and the network delay can be modelled as the finite-state discrete-time Markov chains across the time slots. Given a control policy $\boldsymbol{\Phi}$, the random process $\boldsymbol{\chi}^t$ is thus a controlled Markov chain with the following state transition probability [8], [9],

$$\Pr\{\boldsymbol{\chi}^{t+1} | \boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t)\} = \Pr\{\boldsymbol{Q}^{t+1} | \boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t)\} \Pr\{\boldsymbol{S}^{t+1} | \boldsymbol{S}^t\} \Pr\{\boldsymbol{C}^{t+1} | \boldsymbol{C}^t\} \tag{3}$$

For a controlled Markov chain, the transition probability from a present state $\boldsymbol{\chi}^t$ to the next state $\boldsymbol{\chi}^{t+1}$ depends only on the present state $\boldsymbol{\chi}^t$ and the control policy $\boldsymbol{\Phi}(\chi^t)$ acted on the present state. Taking the discounted expectation with respect to the per-slot utilities $U(\boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t))$ over a sequence of network states $\boldsymbol{\chi}^t$, we can obtain the discounted expected value of the MEC network utility [8],

$$V(\boldsymbol{\chi}, \boldsymbol{\Phi}) = \mathrm{E}\,[\alpha \cdot \sum_{t=1}^{\infty} \gamma^{t-1} U(\boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t)) | \boldsymbol{\chi}^1], \tag{4}$$

where $\alpha, \gamma \in [0, 1)$ are the parameters. γ is a discount factor that discounts the utility rewards received in the future, and $(\gamma)^{t-1}$ denotes the discount to the (t −1)-th power. $\boldsymbol{\chi}^1$ 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 $\boldsymbol{\Phi}$. We let $\alpha = 1 - \gamma$, thus, the expected undiscounted long-term average utility, $\bar{U}(\boldsymbol{\chi}, \boldsymbol{\Phi}) = \mathrm{E}\left[\lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t=1}^{T} U(\boldsymbol{\chi}^t, \boldsymbol{\Phi}(\chi^t)) \middle| \boldsymbol{\chi}^1\right]$ can be considered as a special case of (4) when γ approaches 1 and $\alpha = (1 - \gamma)$ approaches 0 [9]. On the other hand, if γ is set to be 0, then $V(\boldsymbol{\chi}, \boldsymbol{\Phi}) = U(\boldsymbol{\chi}^1, \boldsymbol{\Phi}(\boldsymbol{\chi}^1))$, that is, only the immediate utility performance is considered. We therefore consider the expected discounted long-term utility performance in (4) as a general QoS indicator in this paper.

The objective is to design an optimal task assignment control policy $\boldsymbol{\Phi}^*$ that maximizes the expected discounted long-term utility performance, that is,

$$\boldsymbol{\Phi}^* = arg \max_{\boldsymbol{\Phi}} (V(\boldsymbol{\chi}, \boldsymbol{\Phi})) \tag{5}$$

$V^*(\boldsymbol{\chi}) = V(\boldsymbol{\chi}, \boldsymbol{\Phi}^*)$ is the optimal state value function. The stochastic task assignment optimization in (5) can be considered as a MDP with the discounted utility criterion since the network states follow a controlled Markov process. The optimal task assignment control policy achieving the maximal state value function can thus be obtained by solving the following Bellman's optimality equation [9], [10],

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Task Management for Cooperative Mobile Edge Computing." Presented at Fifth ACM/IEEE Symposium on Edge Computing: Third Workshop on Hot Topics on Web of Things. November 11, 2020 - November 13, 2020.

$$V^*(\boldsymbol{\chi}) = \max_{\boldsymbol{\Phi}}\{(1-\gamma)\, U(\boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})) + \gamma \textstyle\sum_{\boldsymbol{\chi}'} \Pr\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\} V^*(\boldsymbol{\chi}')\}, \quad (6)$$

where $\boldsymbol{\chi}' = (\boldsymbol{Q}', \boldsymbol{S}', \boldsymbol{C}')$ is the MEC network state in the subsequent time slot, and $\Pr\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\}$ represents the state transition probability that making the task assignment $\boldsymbol{\Phi}(\boldsymbol{\chi})$ in state $\boldsymbol{\chi}$ will produce the next 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.

Solving (6) is generally a challenging problem. Traditional approaches are based on value iteration, policy iteration, and dynamic programming [11], [12]. However, these methods require full knowledge of the network state transition probabilities and task arrival statistics that cannot be known beforehand for our problem.

III. Problem Simplification and Online Learning Algorithm

In this section, we focus on developing an algorithm to obtain the optimal task assignment policy with no requirement for prior knowledge of the statistical information about network state transitions and task arrivals by employing online reinforcement learning techniques [13], [14]. However, the task assignment optimization problem in (6) is very complex; both the MEC system state space and the control action space are very large as discussed later. To solve it, first, we simplify the problem by introducing a post-decision state and then reduce the number of system states through decomposition.

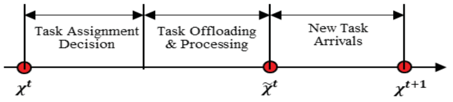


Figure 2. Three phases of a time slot.

Based on the observation that task arrivals are independent of the task assignment policy, we define an intermediate state called post-decision state for each scheduling slot, which is the state after an edge node finishes task offloading to other nodes and local processing. A time slot can be considered consisting of three phases, task assignment decision, task offloading and processing, and new task arrivals as shown in Fig. 2. In phase I, the MEC controller determines the task assignment matrix $\boldsymbol{\Phi}(\boldsymbol{\chi})$ and informs the edge nodes of the task assignment decision. In phase II, an edge node offloads tasks to other nodes or receives tasks from other nodes and processes their tasks based on the task assignment decision. The network state then moves into the post-decision state. The new tasks from the associated devices will arrive at edge nodes in phase III. Note that the three phases and the post-decision state are used to derive the optimal task assignment. In practice, the tasks may arrive at an edge node at any time, and the edge node can process the tasks in its queue and forward the tasks to other nodes during the whole slot time.

At the current scheduling slot, we define the post-decision state as $\tilde{\boldsymbol{\chi}} = (\tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{C}})$, where the node processing and network delay states of the post-decision will remain the same as those at the beginning of the time slot, that is, $\tilde{\boldsymbol{S}} = \{\tilde{s}_n : n \in \mathcal{N} \cup n_c\}$ with $\tilde{s}_n = s_n$ and $\tilde{\boldsymbol{C}} = \{\tilde{\boldsymbol{c}}_n : n \in \mathcal{N} \cup n_c\}$ with $\tilde{\boldsymbol{c}}_n = \boldsymbol{c}_n$, respectively, because they are independent of the task assignment decision. The queue state of post-decision is $\tilde{\boldsymbol{Q}} = \{\tilde{Q}_n : n \in \mathcal{N} \cup n_c\}$ with $\tilde{Q}_n = \max\{Q_n + \Sigma_{i \in e_n} \phi_{i,n} - \Sigma_{i \in e_n} \phi_{n,i} - s_n, 0\}$. The probability of MEC network state transition from $\boldsymbol{\chi}$ to $\boldsymbol{\chi}'$ can then be expressed as,

$$\Pr\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\} = \Pr\{\boldsymbol{\chi}' | \tilde{\boldsymbol{\chi}}\} \Pr\{\tilde{\boldsymbol{\chi}} | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\} = \textstyle\prod_{n,j \in \mathcal{N} \cup n_c} \Pr\{A_n\} \Pr\{s'_n | s_n\} \Pr\{c'_n | c_n\} \quad (7)$$

where $\Pr\{\tilde{\boldsymbol{\chi}} | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\} = 1$ and $A_n = Q'_n - \tilde{Q}_n$. We can control the task assignment decision to ensure that no task drop occurs in the transition to the post-decision state, i.e. the task drop due to buffer overflow may happen only when the new tasks arrive. By introducing the post-decision state, we are able to factor the utility function in (2) into two parts, which correspond to $U_n^{(d)}$ and $U_n^{(o)}$. Then, the optimal state value function satisfying (6) can hence be rewritten by,

$$V^*(\boldsymbol{\chi}) = \max_{\boldsymbol{\Phi}}\{(1-\gamma) \textstyle\sum_{n \in \mathcal{N}} w_d U_n^{(d)}(\boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})) + \tilde{V}^*(\tilde{\boldsymbol{\chi}})\} \quad (8)$$

where $\tilde{V}^*(\tilde{\boldsymbol{\chi}})$ is the optimal post-decision state value function. that satisfies Bellman's optimality equation,

$$\tilde{V}^*(\tilde{\boldsymbol{\chi}}) = (1-\gamma) \textstyle\sum_{n \in \mathcal{N}} w_0 U_n^{(0)}(\boldsymbol{\chi}, \boldsymbol{\Phi}^*(\boldsymbol{\chi})) + \gamma \textstyle\sum_{\boldsymbol{\chi}'} \Pr\{\boldsymbol{\chi}' | \tilde{\boldsymbol{\chi}}\} V^*(\boldsymbol{\chi}') \quad (9)$$

From (8), we find that the optimal state value function can be obtained from the optimal post-decision state value function by performing maximization over all feasible task assignment decisions. The optimal task assignment policy is thus expressed as follows, which should satisfy the maximal delay and task drop constraints.

$$\boldsymbol{\Phi}^* = \operatorname*{argmax}_{\boldsymbol{\Phi}}\{(1-\gamma) \textstyle\sum_{n \in \mathcal{N}} w_d U_n^{(d)}(\boldsymbol{\chi}, \Phi(\boldsymbol{\chi})) + \tilde{V}^*(\tilde{\boldsymbol{\chi}})\} \quad \text{s.t. } d_n \le d^{(\max)} \text{ and } o_n \le o^{(\max)} \quad (10)$$

The task arrival statistics and task processing capability of the edge nodes are independent each other. We can then decompose the optimal post-decision state value function [15]. Mathematically, that is

$$\tilde{V}^*(\tilde{\boldsymbol{\chi}}) = \textstyle\sum_{n \in \mathcal{N}} \tilde{V}_n^*(\tilde{Q}_n, \tilde{s}_n, \tilde{\boldsymbol{c}}_n) \quad (11)$$

Given the optimal control policy $\boldsymbol{\Phi}^*$, according to (9) and (11), the post-decision state value function $\tilde{V}_n^*(\tilde{Q}_n, \tilde{s}_n, \tilde{\boldsymbol{c}}_n)$ satisfies,

$$\tilde{V}_n^*(\tilde{Q}_n, \tilde{s}_n, \tilde{\boldsymbol{c}}_n) = (1-\gamma) w_0 U_n^{(0)*}(Q_n, s_n, \boldsymbol{c}_n) + \gamma \textstyle\sum_{A_n, s'_n, c'_n} \Pr\{A_n\} \Pr\{s'_n | s_n\} \Pr\{\boldsymbol{c}'_n | \boldsymbol{c}_n\} V_n^*(Q'_n, s'_n, \boldsymbol{c}'_n) \quad (12)$$

Based on (8) and (11), the optimal state value function of edge node *n* in the subsequent time slot, $V_n^*(Q'_n, s'_n, \boldsymbol{c}'_n)$ can be expressed as,

$$V_n^*(Q'_n, s'_n, \boldsymbol{c}'_n) = (1-\gamma) w_d U_n^{(d)*}(Q'_n, s'_n, \boldsymbol{c}'_n) + \tilde{V}_n^*(\tilde{Q}'_n, \tilde{s}'_n, \tilde{\boldsymbol{c}}'_n) \quad (13)$$

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Task Management for Cooperative Mobile Edge Computing." Presented at Fifth ACM/IEEE Symposium on Edge Computing: Third Workshop on Hot Topics on Web of Things. November 11, 2020 - November 13, 2020.

where $\tilde{Q}_n', \tilde{s}_n'$ and $\tilde{\boldsymbol{c}}_n'$ are the local post-decision queue, processing, and network delay states for node *n* in the subsequent scheduling slot, respectively.

The linear decomposition of the post-decision state value function proposed above yields two main benefits. First, in order to derive a task assignment policy based on the global MEC network state, $\boldsymbol{\chi} = \{\boldsymbol{\chi}_n : n \in \mathcal{N} \cup n_c\}$ with $\boldsymbol{\chi}_n = (Q_n, s_n, \boldsymbol{c}_n)$ and $\boldsymbol{c}_n = \{c_{n,j}, c_{j,n} \,:\, j \in \mathcal{N} \cup n_c\}$, at least $\prod_{n \in \mathcal{N} \cup n_c} \prod_{j \in \mathcal{N} \cup n_c} (\,|Q_n|\,|s_n||c_{n,j}||c_{j,n}|)$ state values should be kept. Using linear decomposition (11), only $(N + 1)|Q_n||s_n|\ \prod_{j \in \mathcal{N} \cup n_c}(|c_{n,j}||c_{j,n}|)$ values need to be stored, significantly reducing the search space in the task assignment decision making. Second, the problem to solve a complex post-decision Bellman's optimality equation (9) is broken into simpler MDPs. By replacing the post-decision state value function in (10) with (11), we can obtain an optimal task assignment policy $\boldsymbol{\Phi}^*$ under a MEC network state $\boldsymbol{\chi}$.

As discussed before, the number of new task arrivals at the end of a scheduling slot as well as the task processing capability of a node and the states of network delay between the nodes for the next scheduling slot are unknown beforehand. In this case, instead of directly computing the post-decision state value functions in (12), we propose an online reinforcement learning algorithm to learn $\tilde{V}_n^*(\tilde{Q}_n, \tilde{s}_n, \tilde{\boldsymbol{c}}_n), \forall n \in \mathcal{N}$ on the fly. Based on the observations of the network state $\boldsymbol{\chi}_n^t = (Q_n^t, s_{n,}^t \boldsymbol{c}_n^t), \forall n \in \mathcal{N}$, the number of task arrivals $A_n^t, \forall n \in \mathcal{N}$, the decision on the number of tasks locally processed, the number of tasks offloaded to other nodes or received from other nodes, the achieved utility $U_n^{(o)*}(Q_n, s_n, \boldsymbol{c}_n)$ at the current scheduling slot *t*, and the resulting network state $\boldsymbol{\chi}_n^{t+1} = (Q_n^{t+1}, s_{n,}^{t+1} \boldsymbol{c}_n^{t+1})$ at the next slot $t + 1$, the post-decision state value function for node *n* can be updated by,

$$\tilde{V}_n^{t+1}(\tilde{Q}_n^t, \tilde{s}_n^t, \tilde{\boldsymbol{c}}_n^t) = (1 - \varepsilon^t)\tilde{V}_n^t(\tilde{Q}_n^t, \tilde{s}_n^t, \tilde{\boldsymbol{c}}_n^t) + \varepsilon^t[(1 - \gamma)w_0 U_n^{(0)}(Q_n^t, s_n^t, \boldsymbol{c}_n^t) + \gamma V_n^t(Q_n^{t+1}, s_n^{t+1}, \boldsymbol{c}_n^{t+1})] \tag{14}$$

where $\varepsilon^t \in [0, 1)$ is the learning rate. The task assignment matrix $\boldsymbol{\Phi}^t = [\phi_{n,j}^t : n, j \in \mathcal{N} \cup n_c\}$ at scheduling slot *t* is determined as,

$$\boldsymbol{\Phi}^t = \underset{\boldsymbol{\Phi}}{\operatorname{argmax}}\{\sum_{n \in \mathcal{N}} [(1 - \gamma)w_d U_n^{(d)}(Q_n^t, s_n^t, \boldsymbol{c}_n^t) + \tilde{V}_n^t(\tilde{Q}_n^t, \tilde{s}_n^t, \tilde{\boldsymbol{c}}_n^t)]\} \quad \text{s.t. } d_n^t \le d^{(\max)} \text{ and } o_n^t \le o^{(\max)} \tag{15}$$

The state value function of node *n* at slot $t + 1$ is evaluated by,

$$V_n^t(Q_n^{t+1}, s_n^{t+1}, \boldsymbol{c}_n^{t+1}) = (1 - \gamma)w_d U_n^{(d)}(Q_n^{t+1}, s_n^{t+1}, \boldsymbol{c}_n^{t+1}) + \tilde{V}_n^t(\tilde{Q}_n^{t+1}, \tilde{s}_n^{t+1}, \tilde{\boldsymbol{c}}_n^{t+1}) \tag{16}$$

The online learning algorithm for estimating the optimal post-decision state value function and determining the optimal task assignment policy is summarized in Algorithm 1.

Algorithm 1. Online Learning Algorithm for Optimal Post-Decision State Value Function

```
1.  Initialize the post-decision state value functions $\tilde{V}_n^t(\tilde{\chi}_n^t)$, $\forall \tilde{\chi}_n^t$ and $\forall n \in \mathcal{N}$ for $t = 1$.
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 determines the task assignment matrix, $\boldsymbol{\Phi}^t = [\boldsymbol{\phi}_n^t : n \in \mathcal{N}]$ according to (15).
3.  After offloading and processing the tasks according to the above task assignment decision, the controller observes the post-decision state, $\tilde{\boldsymbol{\chi}}^t = \{\tilde{\boldsymbol{\chi}}_n^t : n \in \mathcal{N}\}$, where $\tilde{\boldsymbol{\chi}}_n^t = (\tilde{Q}_n^t, \tilde{s}_n^t, \tilde{\boldsymbol{c}}_n^t)$ with $\tilde{Q}_n^t = \max\{Q_n^t + \Sigma_{i \in e_n}\phi_{i,n}^t - \Sigma_{i \in e_n}\phi_{n,i}^t - s_n^t, 0\}$, $\tilde{s}_n^t = s_n^t$, and $\tilde{\boldsymbol{c}}_n^t = \boldsymbol{c}_n^t$.
4.  With $\boldsymbol{A}^t = \{A_n^t : n \in \mathcal{N}\}$ new tasks arrived at the end of slot $t$, the network state transits to $\boldsymbol{\chi}^{t+1} = \{\boldsymbol{\chi}_n^{t+1} : n \in \mathcal{N}\}$ where $\boldsymbol{\chi}_n^{t+1} = (\tilde{Q}_n^t + A_n^t, s_n^{t+1}, \boldsymbol{c}_n^{t+1})$ at the following scheduling slot $t$+1.
5.  Calculate $V_n^t(Q_n^{t+1}, s_n^{t+1}, \boldsymbol{c}_n^{t+1})$, $\forall n \in \mathcal{N}$ according to (16) and updates the post-decision state value functions $\tilde{V}_n^{t+1}(\tilde{Q}_n^t, \tilde{s}_n^t, \tilde{\boldsymbol{c}}_n^t)$, $\forall n \in \mathcal{N}$ according to (14).
6.  The scheduling slot index is updated by $t \leftarrow t + 1$.
7   Repeat from step 2 to 6.
```

IV. Numerical Results

We provide the evaluation results in this section and compare the performance of our online reinforcement learning scheme with several benchmark schemes including i) no cooperation, i.e. an edge node processes all the tasks it receives from its associated devices by itself; ii) cloud execution, i.e. an edge node offloads all its received tasks to the cloud data center for execution; iii) one-shot deterministic optimization which is similar to the scheme in [5].

We simulated multiple MEC network scenarios with different system parameters. Due to the page limit, we present the results for a typical setting. We assume the slot duration is 30 ms. The task processing capability of an edge node is considered to be an independent Markov chain model with three states {4, 2, 1} tasks per slot. The network delay between two edge nodes is also modeled as a Markov chain with three states, {1, 0.5, 0.2} slots. The cloud data center has powerful computation resources, and the queuing and processing delay in the cloud data center is small enough to be ignored, but forwarding the tasks to the cloud incurs a large network delay, 10 slots, due to a long distance with many hops over the Internet.

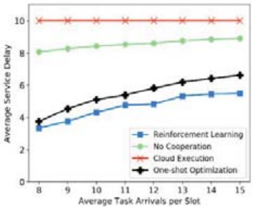


Figure 3. The average task service delay versus the average task arrivals per slot for different algorithms.

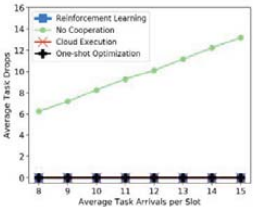


Figure 4. The average number of dropped tasks per slot versus the average task arrivals per slot for different algorithms.

In Fig. 3, we compare the average task service delay for different algorithms with three edge nodes and one cloud data center when the task arrivals follow independent Poisson arrival process and the average number of task arrivals per slot

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Task Management for Cooperative Mobile Edge Computing." Presented at Fifth ACM/IEEE Symposium on Edge Computing: Third Workshop on Hot Topics on Web of Things. November 11, 2020 - November 13, 2020.

at an edge node changes. Note that the delay is measured in the unit of the time slot duration. The curves indicate that our proposed online reinforcement learning scheme outperforms all the three benchmark schemes. Compared to the one-shot optimization algorithm (the second best in terms of the service delay), the proposed online learning scheme can capture the dynamic MEC network state transitions and determine the optimal task assignment matrix by taking into consideration the impacts of time-varying stochastic network environments and node computing capabilities to the expected long-term performance in the future. However, the one-shot deterministic optimization algorithm makes shortsighted task assignment decisions and may cause a lot of tasks to be shipped to the cloud data center for processing, which leads to a large network delay and thus a large service delay. In Fig. 4, we present the average number of the tasks dropped per slot for different algorithms. The task drops for the online learning and one-shot optimization algorithms are close to zero because the algorithms minimize the task drops and the edge nodes will forward the tasks to the cloud data center when their buffers become full. For the no cooperation scheme, when the workload is high, an edge node does not have enough resources to process all the tasks so that the service delay increases and the tasks are dropped. For the cloud execution scheme, there is always a large value of network delay to ship the tasks to the cloud data center for processing over the Internet.

V. Conclusions

In many MEC scenarios, the task arrival statistics, task processing capability at an edge node, and network delay between two nodes are time-varying and unknown beforehand. Therefore, casting the task assignment as a dynamic and stochastic optimization problem is more reasonable and compelling. In this paper, we have proposed and investigated a stochastic framework to model the interactions among various entities of a MEC system, including the edge-to-edge horizontal cooperation and the edge-to-cloud vertical cooperation for jointly processing tasks under dynamic and uncertain network environments. The task assignment optimization problem is formulated as a Markov decision process by taking into consideration the non-stationary computing and network states as well as the interaction and heterogeneity of the involved entities. To solve the problem, we derive an algorithm based on online reinforcement learning, which learns on the fly the optimal task assignment policy without prior knowledge of task arrival and network statistics. Further, considering the structure of the underlying problem, we introduce a post-decision state and a function decomposition technique to reduce the search space, which are combined with reinforcement learning. The evaluation results show that the proposed online learning-based scheme reduces the service delay, compared to the existing schemes that do not consider dynamic changes in traffic and MEC network statistics.

Acknowledgment

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] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, “Computation offloading and resource allocation in wireless cellular networks with mobile edge computing,” *IEEE Trans. Wireless Commun*., vol. 16, no.8, pp. 4924–4938, Aug. 2017.

[2] Y. Mao, J. Zhang, and K. B. Letaief, “Dynamic computation offloading for mobile-edge computing with energy harvesting devices,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[3] C. Do, N. Tran, C. Pham, M. Alam, J. H. Son, and C. S. Hong, “A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing,” in *Proc. of the IEEE ICOIN*, pp. 324–329, Siem Reap, Cambodia, Jan. 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] 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.

[7] X. Chen, Z. Han, H. Zhang, G. Xue, Y. Xiao, and M. Bennis, “Wireless resource scheduling in virtualized radio access networks using stochastic learning*,” IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 961-974, 2018.

[8] S. M. Ross, Introduction to stochastic dynamic programming. Academic press, 2014.

[9] R. Howard, Dynamic Programming and Markov Processes. MIT Press, 1960.

[10] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, Belmont, MA, 1995.

[11] 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.

[12] D. Adelman and A. J. Mersereau, “Relaxations of weakly coupled stochastic dynamic programs,” *Oper. Res.*, vol. 56, no. 3, pp. 712–727, Jan. 2008.

[13] X. Chen, P. Liu, H. Liu, C. Wu, Y. Ji, “Multipath Transmission Scheduling in Millimeter Wave Cloud Radio Access Networks,” in *Proceedings of IEEE ICC'18*, Kansas City, MO, May 2018.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[15] 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.

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. ”Task Management for Cooperative Mobile Edge Computing.” Presented at Fifth ACM/IEEE Symposium on Edge Computing: Third Workshop on Hot Topics on Web of Things. November 11, 2020 - November 13, 2020.

'Passwords Keep Me Safe' – Understanding What Children Think about Passwords

Mary Theofanos, *National Institute of Standards and Technology*
Yee-Yin Choong, *National Institute of Standards and Technology*
Olivia Murphy, *University of Maryland, College Park*

Abstract

Children use technology from a very young age, and often have to authenticate. The goal of this study is to explore children's practices, perceptions, and knowledge regarding passwords. Given the limited work to date and the fact that the world's cyber posture and culture will be dependent on today's youth, it is imperative to conduct cybersecurity research with children. We conducted the first large-scale survey of 1,505 3rd to 12th graders from schools across the United States. Not surprisingly, children have fewer passwords than adults. We found that children have complicated relationships with passwords: on one hand, their perceptions about passwords and statements about password behavior are appropriate; on the other hand, however, they simultaneously do not tend to make strong passwords, and practice bad password behavior such as sharing passwords with friends. We conclude with a call for cybersecurity education to bridge the gap between students' password knowledge with their password behavior, while continuing to provide and promote security understandings.

1 Introduction

School children are engaged in technology and cyber learning at very young ages. In fact, today's primary and secondary school children referred to as "digital natives" [32] or "neo-digital natives" [29] have never experienced a world without technology. Computer technology is just a part of their lives. As a result, children are exposed to more and more systems designed specifically for them as well as accessing and using ubiquitous applications such as social media. Many of these systems require authentication to retain a history of interaction, or to ensure that it is genuinely the child using the system. Without evidence of clearly superior and appropriate alternatives, it is understandable that developers implement passwords. As a result, children are actively and frequently using passwords, making understanding their password practices and behavior important.

Usability testing with children is constrained by strict ethical requirements which may discourage researchers from testing authentication mechanisms with this target group altogether [16, 26]. Most of the research in usable security has focused on adults. Yet, over the next 10 to 20 years the world's cyber posture and culture will be dependent on the cybersecurity and privacy knowledge and practices of today's youth. Without an understanding of extant behavior, it is infeasible to start seeking an alternative, more appropriate, mechanism for child-tailored authentication. Despite extensive studies of password practices of participants over 18 years old (e.g., [1, 7, 14, 17, 31, 43]), children's password practices have not been well studied.

To understand current children's password perceptions and behavior, we conducted a study to answer the following research questions (RQ):

RQ1. Password Understandings:
(a) What do students know about passwords?
(b) Why do they think they need passwords?
(c) What are students' passwords perceptions?

RQ2. Password Behaviors:
(a) How do students create and maintain passwords?
(b) What are the characteristics of passwords they create?

The contributions of this paper are threefold:

1) Firstly, we conducted the first large-scale study on the use, perceptions and behavior of passwords of the United States (US) youth 3rd to 12th grades–Generation Z (Gen Z) those born from the mid-1990's to the late 2000's [29];
2) Secondly, we characterize the state of children's perceptions and knowledge of passwords;
3) Finally, we offer concrete suggestions for next steps in both youth password research and education.

We next review related work. We present our methodology followed by results, discussion and conclusions.

2 Related Research

In 2015, 94% of US children between the ages of 3 and 18 had a computer at home, and 86% of children had internet access at home [39]. As of 2019, 53% of children own their own smartphone by age 11, with that number rising to 84% among teenagers [11]. Children around the world are going online more, at younger ages, and in more diverse ways [13]. Children spend more time on screen media performing

Theofanos, Mary; Choong, Yee-Yin. "'Passwords Keep Me Safe' – Understanding What Children Think about Passwords." Presented at 30th USENIX Security Symposium. August 11, 2021 - August 13, 2021.

various activities such as TV/videos, gaming, browsing websites, and social media [11]. As children are doing more activities online, they are creating user accounts and passwords as required by those online systems. However, the research topic on children's password perceptions and practices has not been extensively studied, so there is a comparative lack of literature available.

In 2019, Choong *et al* [9] performed a systematic search on cybersecurity research involving children and classified 78 papers into two major categories – Designing for Children, and Children & Authentication which each was further broken into six sub-categories. They identified a gap in the literature related to children's password comprehension and practices. This present study seeks to fill that gap.

Several researchers performed empirical studies on children's passwords with small numbers of participants, usually with narrow (two years) age ranges (e.g., [21, 27, 33]). These studies agree that the younger a child is the less complex their passwords are and should be required to be due to age-specific factors like memory and spelling, and that children frequently use personal information in password creation [21, 27, 33]. Other researchers used surveys to gather larger amounts of data on children's password knowledge and behaviors and found similar results. For example, Rim and Choi [35] analyzed password generation types from 550 middle and high school students in South Korea and concluded that students are likely to use personal information in their passwords. Further, the study found that participants seldom worried about protecting passwords and personal information. This is concerning because, as revealed in Irwin's [23] investigation of 258 10th to 12th grade South African Students' risk taking behavior and awareness, students in this age group have a high level of risk and gaps in their risk awareness and avoidance behavior. Coggins [10] conducted a small-scale survey on children's password knowledge from 74 4th to 6th grade students that supports all of the above studies, finding that 70% of participating students used personal information in their passwords and 32% had experienced hacking. Our present study seeks to build upon these findings by investigating a full range of school-age students from 3rd to 12th grade, and exploring not only students' password behavior, but also their perceptions and understandings about the role of passwords.

In addition to the field of knowledge surrounding children's password behavior, several studies have investigated children's perceptions of online privacy and security more broadly. For example, Kumar *et al* [24] interviewed 18 US families with children ages 5 to 11, and found that children on the upper end of that age range generally recognized certain privacy and security components, but that younger participants (5-7) had gaps in their knowledge. Zhang-Kennedy *et al* [45] similarly conducted interviews with 14 Canadian parent-child dyads with children ages 7 to 11 to understand their concept of privacy and perceptions of online threats. The study found that children and adults view online privacy and security differently, with children being less concerned than their parents about security threats and mostly worried about threats from local (family, friends, etc.) sources. Our present study seeks to combine the focus on perception in the above studies with an emphasis on password knowledge and understandings as well as password use.

Methodologically speaking, researchers frequently use surveys and questionnaires in order to understand children's perceptions and awareness of online safety, privacy and security. For example, Žufić *et al* [46] administered three surveys over the course of eight years to 1,232 students ages 7 to 15 in Croatia to find that student use of information-telecommunication technology is increasing over time, but student safety awareness is not. Yilmaz *et al* [44] similarly deployed a survey to 2,029 Turkish high school students and revealed that only about half of the students surveyed have high awareness of how to ensure information security toward threats. Paluckaitė *et al* [30] survey of 152 Lithuanian adolescents' perceptions of risky online behavior adds nuance to these security threat understandings by revealing that many participants do understand risky behavior as risky but still engage in them, which may or may not be a product of their awareness of privacy and security threats. Across the board, these studies serve as precedents for our own use of surveys to investigate students' password use, perceptions, and behaviors.

Based on the literature reviewed above, currently existing research often uses a small sample size, does not cover a full age range of K-12 students, and usually does not offer inferential comparisons among kids at different developmental stages in order to gain insight on age-related progression in children's understanding of cybersecurity and privacy. While there have been a few larger-scale survey studies, they have been all focusing on children outside of the US. Investigation in this area to understand and gauge current levels of US children's comprehension and practice related to passwords is essential to provide insights into overall children's cybersecurity hygiene. This study seeks to add to the burgeoning field of scholarship surrounding children's password use, perceptions, and understandings while also addressing the aforementioned shortcomings in the field by conducting a large-scale survey of students between ages 8 and 18 (3rd to 12th grades) in the United States.

3 Method

We developed a large-scale, self-report survey to understand what challenges US grade school children face regarding passwords. The target population was students from 3rd to

12th grades (ages of 8 to 18 years old). The goal was to identify students' practices, perceptions, and knowledge regarding passwords. Each student answered questions assessing their use of computers, passwords, password practices, knowledge about and feelings about passwords, together with information about grade and gender.

3.1 Survey Development

The research questions guided the development of survey objectives for accessing student's use of computers, of passwords, password practices, knowledge about passwords, feelings about passwords, and tests for age differences. A list of possible items was generated targeting the objectives. All of the items were closed response except for two numerical response and two open response items where students were asked: how many passwords they have; how many times a day they use passwords; to list a reason(s) why people should use passwords, and to create a new password for a given scenario.

Early in survey development, feedback from teachers and a pilot survey suggested that two surveys featuring the same questions but using different, age-appropriate language would be required to accommodate the wide age range of the intended student population. Thus, two surveys were designed: a 15-item survey for 3rd to 5th graders, and a 16-item survey for 6th to 12th graders. The extra item in the 6th to 12th grade survey asked students whether they have experience helping their family members with passwords. The content of the other 15 questions was identical across the two surveys, with the language and format of the response variables adjusted to be age appropriate. For example, most of the response variables were "*Yes*" or "*No*" for the 3rd to 5th graders, while the 6th to 12*th* graders' response variables were more detailed and they were asked to check all variables that apply.

To ascertain the content and construct validity of the survey instruments, four types of reviews were conducted iteratively. Content experts in usable security were asked to evaluate the alignment matrix and provide feedback on the alignment of the categories with the scope of the survey goals, the alignment of the items with the category, and the possibility of missing items. Survey experts also reviewed each item for clarity for the intended audience, appropriate format, and alignment of response options. Content experts (elementary, middle and high school teachers) focused on the language and format of the items based on the grade/age of the students. As a pilot, cognitive interviews with students were also conducted using a talk-aloud protocol to determine if the questions were being appropriately interpreted. Cognitive probing techniques where students were asked to both paraphrase items (e.g., "*How would you ask the question in your own words*") and interpret them (e.g., "*What is your answer and why*") complemented the talk-aloud protocol. After each type of review, the survey instruments were refined based on the feedback and comments. The final surveys were converted to Scantron© forms–machine readable paper forms as shown in the Appendix.

3.2 Procedure & Recruitment

The National Institute of Standards and Technology Institutional Review Board reviewed and approved the protocol for this project and all subjects provided informed consent in accordance with 15 CFR 27, the Common Rule for the Protection of Human Subjects. The sampling plan focused on recruiting participants from at least three different school districts from three different US regions–the East, South, and Midwest–in order to collect a geographically diverse and more nationally representative sample population. Principals and teachers from the selected districts were recruited using a snowball sampling approach. The principals were to determine which classrooms would participate, and the selected classroom teachers would distribute parental consent forms.

The schools, individual teachers, and students that participated were compensated. Each school received $1000, the teachers received $50 gift cards, and the students received age-appropriate trinkets such as caricature erasers or ear buds, for example. Each participating classroom also received $50 for a classroom thank-you celebration where all students celebrated. Parental consent and student assent forms were collected prior to survey distribution. The survey administration was tailored for the appropriate age group: all children completed Scantron© survey forms, with teachers reading the survey aloud in the 3rd to 5th grades. The data were collected anonymously. All open-ended responses were manually entered into a spreadsheet by the researchers. Each completed survey was assigned a unique random participant identifier, for example, P1234.

3.3 Participants

A total of 1,505 3rd to 12th grade students from schools across the South, Midwest, and Eastern regions in the United States completed the survey. Demographics are shown in Table 1.

Students	#	**Gender** (%)			**Age** (Years)	
		Boy	Girl	Others[1]	Mean	SD
ES	425	40.2	51.9	7.9	9.03	0.92
MS	357	45.1	50.3	4.6	12.46	1.01
HS	723	44.7	51.4	3.9	15.79	1.21

Table 1. Participant Demographics

[1] This includes "other" and "prefer not to answer" responses.

Participants included 425 3rd to 5th grade elementary school students (ES) from four elementary schools, 357 6th to 8th grade middle-school students (MS) from four middle schools, and 723 9th to 12th grade high school students (HS) from three high schools.

3.4 Data Analysis Procedure

Descriptive statistics were used to report the frequency and percentage of the categories that participants chose as responses to the multiple-choice questions. We compared groups using inferential statistics with an overall significance level set at $\alpha = 0.05$.

For categorical variables, Chi-Square tests of association were used, with effect size calculated using Cramer's V. For measured variables with interval levels, data were first tested for normality. Nonparametric tests (Mann-Whitney U test to compare two groups) were applied as the data were not normally distributed. P*ost-hoc* comparisons were used to compare groups: ES vs. MS, MS vs. HS, and ES vs. HS while applying the Holm-Bonferroni method to control the family-wise error rate [19] with adjusted $\alpha = 0.017$.

Qualitative responses to the open-ended question "*Why do you think people should use passwords?*" were coded using a two-cycle coding process [36]. In the first cycle, inductive thematic and in vivo coding were used separately by two members of the research team, and then discussed and merged into one set of codes and sub-codes. We calculated intercoder reliability for the initial coding of the data using the ReCal2[2] software, the Krippendorf's Alpha score was 0.968. Second cycle pattern coding was used to condense the larger code deck into major themes, and returned three final thematic codes–access, privacy, and safety–that were applied to all of the data [36]. A third, qualitatively trained researcher was then brought in to independently conduct the same inductive two-cycle coding process to further validate results, and to advise on qualitative thematic consolidation and discussion. The third coder returned four themes: safety, privacy, offensive and defensive access, and protection. The new theme "protection" was discussed by the research team and also applied to the data.

The third researcher also performed a single-cycle deductive thematic coding of the responses to the second open-response survey question asking participants to create a password. The themes for the deductive coding–perceived personal information, number or word-only, alphanumeric, and strong/weak–were derived from the afore cited literature in order to check the validity of collected data with currently existing theories and research surrounding children's password creation behavior.

Any quotes provided within this paper as exemplars are verbatim from the children's responses. The quotes are presented in italics and followed by a notation with the unique participant identifier and the participant's grade. For example, (P745, 3rd) indicates a quote from P745 who was a 3rd grade student.

4 Results

As indicated in section 3.4, the significance level of statistical analyses was set at $\alpha = 0.05$ and adjusted $\alpha = 0.017$. The asterisk symbol "*" is used to indicate statistical significance ($p < \alpha$).

4.1 Current Usage

To understand our participants' current usage of computing devices, we collected data on the types of devices as well as activities performed with those devices. The percentages of computing device usage are summarized in Table 2. When comparing among ES, MS, and HS, the MS reported using laptop the least, followed by ES, then HS ($\chi^2 = 43.83$, df = 2). The use of tablets decreases significantly from ES to MS, to HS ($\chi^2 = 46.17$, df = 2), whereas cell phone usage increases significantly from ES to MS, to HS ($\chi^2 = 180.65$, df = 2).

Grade	Desktop (%)	Laptop* (%)	Tablet* (%)	Cell phone* (%)	Gaming console (%)
ES	74.57	84.07	71.86	63.22	68.86
MS	63.28	74.01	53.95	84.75	66.38
HS	61.91	89.20	46.68	91.41	55.68

Table 2. "*What types of computers do you use at school and at home?*"

Students use computers for many activities such as schoolwork, homework, games, texting, and social media (Table 3).

Response Option	ES (%)	MS (%)	HS (%)
Email*	28.15	25.71	57.62
Entertainment	87.90	81.92	82.27
Games*	92.95	77.12	63.85
Homework*	59.59	59.60	86.98
Internet	84.58	73.45	82.69
School	83.50	71.47	87.95
Social media*	38.22	57.91	71.88
Texting*	46.30	55.08	70.36

Table 3. "*What do you do on computers?*"

HS significantly do more homework compared to ES ($\chi^2 = 151.99$, df = 1) and compared to MS ($\chi^2 = 106.22$, df = 1). HS also use emails significantly more than ES ($\chi^2 = 116.40$, df = 1) and more than MS ($\chi^2 = 98.55$, df = 1). When comparing

[2] http://dfreelon.org/utils/recalfront/recal2/

among ES, MS, and HS, social media use increases significantly from ES to MS, to HS ($\chi^2 = 153.79$, df = 2). Likewise, texting increases significantly from ES to MS, to HS ($\chi^2 = 95.83$, df = 2). Finally, playing games decreases significantly from ES to MS, to HS ($\chi^2 = 75.14$, df = 2).

4.2 Password Understandings

Students reported learning about good password practice mainly from home (72.35%) and school (59.90%) as opposed to learning from internet (24.48%) and friends (12.28%).

4.2.1 Why Passwords?

Students were asked "*Why do you think people should use passwords?*" ES were asked to provide one reason while MS and HS were asked to provide up to three reasons.

As mentioned previously, the responses were coded using a two-cycle thematic process. There were 7 primary codes/sub-codes and 20 in vivo operationalization terms for those codes, such as "security." The final code book of primary codes, sub-codes, and in vivo terms is shown in Table 4.

Primary Code	Sub-code	Code Operationalization
Access		Mentioned the ability (i.e., allow access) or inability (i.e., prevent access) to use accounts, devices, data, information
	Hacking	Mentioned *hack* or *hacking* (literally), or *scam*
Privacy		Mentioned *private*, *privacy*, *confidentiality*, or *secret* (literally)
Protection		Mentioned *protect* or *protection* (literally); to avoid loss (such as data/information, devices, finances/money); concerned with personal or physical protection
Safety		Mentioned *safe* or *safety* (literally), or mentioned *track*(*ing*), *stalk*(*ing*), *cyberbully*, or *kidnap*; concerned with online harm from bad people; concerned with personal or physical safety
	Security	Mentioned *secure* or *security* (literally)
	Steal	Mentioned *steal*, *stolen*, or *theft* (literally)

Table 4. Why Passwords – Qualitative Analysis Code Book

The percentages of responses in each primary and sub-code are shown in Table 5. As shown in Table 5, for ES, *Access* was the most frequently provided reason for passwords for ES, followed by *Safety*. The ES' responses included both preventing access and providing access. Response examples were "*To keep people out of their stuff*" (P745, 3rd) and "*They should use it because the computer needs to know who they are"* (P623, 5th). Representative examples for *Safety* included "*To keep us safe*" (P1131, 4th), "*To keep their stuff safe*" (P722, 5th) and "*... because someone might track you down*" (P691, 3rd). Almost all MS cited *Access*, but *Privacy* was the second most common response. Exemplar MS' responses include *Access*: "*To lock up everything"(*P2652, 7th) and "*So people don't login and be nos[e]y*" (P1665, 8th); *Privacy*: "*To keep their information private*" (P2909, 6th) and "*To keep stuff private*" (P2918, 8th). HS were focused on *Privacy* followed by *Access*. Representative HS' responses include: *Privacy*: "*Keep things private*" (P1768, 10th) and "*To keep privacy*" (P2596, 12th); *Access*: "*So no one will get in your stuff"(*P2007, 9th) and "*To keep unwanted people off your device*" (P1392, 11th).

Primary Code	Sub-code	**ES** (%)	**MS** (%)	**HS** (%)
Access		43.04	100.58[3]	61.52
	Hacking	11.14	19.31	11.38
Privacy		19.49	52.16	71.07
Protection		2.78	22.48	31.32
Safety		26.84	39.19	34.27
	Security	0.76	8.65	27.95
	Steal	3.54	12.68	5.62

Table 5. Children's Responses to Why We Need Passwords

Protection, *Security*, *Hacking*, and *Steal* are the remaining codes/sub-codes. *Protection* was cited more frequently by HS and MS than ES. Examples include: "*To be protected*" (P2893, 6th) and "*To protect information*" (P2719, 12th). *Security* was reported more by HS than MS and ES. Example responses include: "*Security reasons*" (P244, 9th) or "*Keep info secure"* (P1319, 12th). *Hacking* was mentioned more frequently by MS, for example, "*to make it harder to get hacked*" (P1433, 6th). *Steal* received the fewest responses across all three age groups (13 % and below). Responses such as "*So people won't steal your account*" (P2968, 8th) and "*if someone steals your phone*" (P2940, 7th) were common themes in the *Steal* coded data.

[3] Note: a single student's responses can be coded to multiple sub-codes that belong to the same primary code which may result in percentages over 100 %, for example, *Access* for MS.

4.2.2 *Password-Related Perceptions*

In general, over 50% of the students found it easy to make a password, but less than 50 % found it easy to make many different passwords (Figure 1).

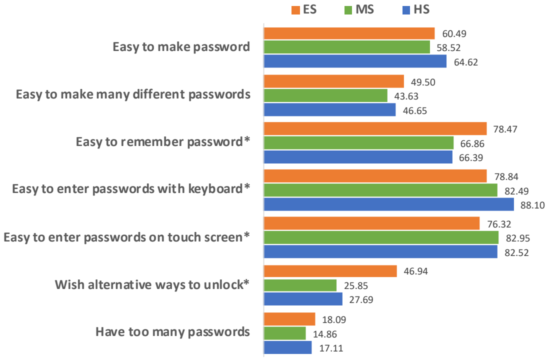


Figure 1. Children's Perception of Passwords (in %)

ES found it significantly easier to remember passwords, compared to MS ($\chi^2 = 6.74$, df = 1) and compared to HS ($\chi^2 = 9.60$, df = 1). While generally students reported it easy to enter passwords (more than 75%) with keyboard or on touch screen, there were significant differences when comparing ES to their older counterparts. Entering password with keyboard becomes significantly easier from ES, to MS, then to HS ($\chi^2 = 32.33$, df = 2). ES found it significantly more difficult to enter passwords on touch screens compared to MS ($\chi^2 = 11.75$, df = 1) and HS ($\chi^2 = 16.47$, df = 1). Finally, significantly more ES wanted alternative ways (other than passwords) to authenticate compared to MS ($\chi^2 = 32.56$, df = 1) and to HS ($\chi^2 = 37.77$, df = 1). Across all three age groups, less than 20 % reported having too many passwords.

4.3 Password Behaviors

4.3.1 *Password Habits*

Children's password habits are summarized in Table 6.

Response Option	ES (%)	MS (%)	HS (%)
Change passwords*	61.08	78.06	74.13
Keep passwords private*	92.96	97.71	98.46
Share passwords with friends*	22.66	39.49	44.71
Sign out after use	92.07	96.57	92.29
Use the same password for everything*	57.82	80.63	87.29

Table 6. Children's Password Habits

While more than 92% of each group reported that they keep their passwords private, ES reported significantly lower percentage compared to MS ($\chi^2 = 18.18$, df = 1) and to HS ($\chi^2 = 47.21$, df = 1). However, as children age from ES to MS, to HS, they progressively reported significantly more and more that they "share passwords with friends" ($\chi^2 = 60.68$, df = 2). The use of same password for everything also increases significantly from ES, to MS, to HS ($\chi^2 = 149.02$, df = 2). ES reported "change passwords" significantly less often compared to MS ($\chi^2 = 29.59$, df = 1) and to HS ($\chi^2 = 29.06$, df = 1). The two primary reasons (over 60 %) for changing passwords are "*when I forgot my passwords*" and "*when someone finds out my passwords.*" All age groups reported a very high rate (more than 92%) of signing out after use.

4.3.2 *Password Selection & Storage*

When asked how they get their passwords, all are given passwords by their schools at very high rates as over 80% as summarized in Table 7.

Response Option	ES (%)	MS (%)	HS (%)
Given by School	88.83	82.39	87.79
Make my own passwords*	54.50	81.53	95.28
Made by parents*	45.69	19.60	7.07
Made my own with parents' help*	44.25	17.90	8.32

Table 7. "*How do you get your passwords?*"

As shown in Table 7, younger students (ES) reported having significantly more parental involvement in creating their passwords. Students having passwords made by parents decrease significantly from ES to MS, to HS ($\chi^2 = 209.07$, df = 2). Similarly, students making their own passwords with parents' help decrease significantly from ES to MS, to HS ($\chi^2 = 179.13$, df = 2). And, students making their own passwords increase significantly from ES to MS, to HS ($\chi^2 = 311.09$, df = 2).

Figure 2 shows how students remember passwords. More than 89 % of participants across age groups reported memorizing their passwords as a strategy for remembering passwords.

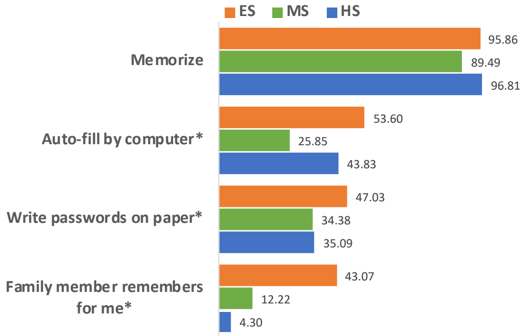


Figure 2. "*How do you remember your passwords?*" (in %)

Approximately half of ES reported that they write their passwords on paper which was significantly higher than MS ($\chi^2 = 9.47$, df = 1) and HS ($\chi^2 = 10.66$, df = 1). The MS reported using auto-fill feature less frequently compared to

ES ($\chi^2 = 52.22$, df = 1) and compared to HS ($\chi^2 = 33.77$, df = 1). As children age, their relying on family members to remember their passwords significantly decreases from ES to MS, to HS ($\chi^2 = 267.96$, df = 2).

Both MS and HS were asked an additional question on whether they help their family members with passwords. About 47 % of MS and 34 % of HS chose "*Yes*." Of those who chose "*Yes*," the primary assistance they provided was to "*Help family members remember passwords*"–MS (68.86 %) and HS (78.01 %).

4.3.3 Created Password Analysis

The three groups were asked to create a password: "*Let's say you just got a new game to play on the computer, but you need a password to use it. Please make up a new password for that game. (Remember, don't write down one of your real passwords.)*"

Password Characteristics

On average, students created passwords about 10 characters long (ES: 9.90 characters, MS: 10.42 characters, and HS: 10.44 characters). Using the *Mann-Whitney U* test , ES was found creating significantly shorter passwords, compared to MS (z = -3.23) and HS (z = -4.75).

Figure 3 shows the distribution of different character types used in the passwords created by the participants. Lowercase letters make up the majority of the passwords, followed by numbers. ES used significantly fewer lowercase letters, compared to MS (z = -3.44) and HS (z = -5.42). ES used significantly more numbers than MS (z = 2.52) and HS (z = 2.40). Across all age groups, symbols or white spaces were rarely used.

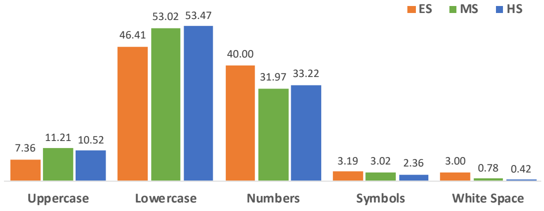


Figure 3. Character Types in Passwords (in %)

We further examined character type positioning in the passwords. Figures 4, 5, and 6 display the overall character type distributions relative to their positions in the passwords, for password lengths of 9 (median) for ES, and password lengths of 10 (median) for MS and HS.

As shown in Figure 4, ES predominantly used lowercase letters and numbers. They tend to start their passwords with numbers or uppercase letters in the 1st position. Immediately after the 1st position, the remaining positions, lowercase letters were used predominantly (about 50 %) and numbers were used between 39 % and 46 %.

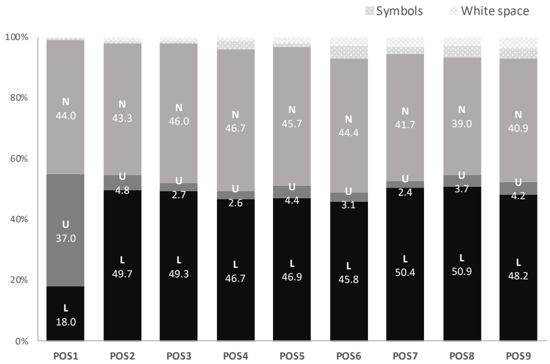


Figure 4. Character Types by Positions in Passwords (ES)
(in %; L – lowercase, U – uppercase, N – numbers)

In contrast, the patterns for MS (Figure 5) and HS (Figure 6) look quite different from ES. Both MS and HS also tend to start their passwords with uppercase letters (about 55%), but numbers are not as prevalent in the first position as for ES. We observe a decreasing use of lowercase and increasing trend of using numbers as the position gets higher.

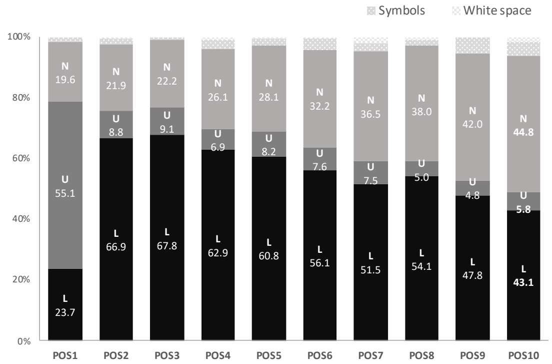


Figure 5. Character Types by Positions in Passwords (MS)
(in %; L – lowercase, U – uppercase, N – numbers)

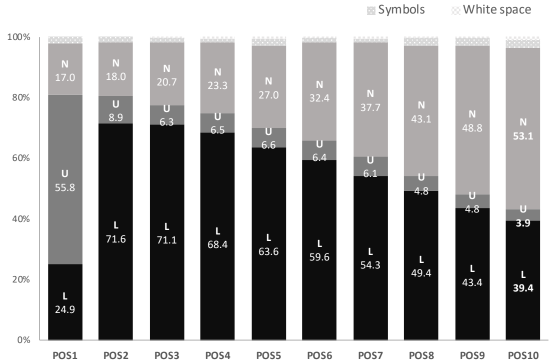


Figure 6. Character Types by Positions in Passwords (HS)
(in %; L – lowercase, U – uppercase, N – numbers)

In addition, the passwords did not use a broad range of characters, much like adults [22]. For all three age groups, only 8 alphabetic characters and four numbers "0, 1, 2, 3" were used with frequency higher than or equal to 3 %.

Many of the passwords contained passphrases or multiple common words. We specifically examined the passwords for the following three characteristics (Table 8):

- *Dictionary word*: a single dictionary word,
- *Dictionary word plus*: a single dictionary word plus numbers and special characters preceding or following the word,
- *Numbers only*: passwords contain all numbers.

Password Characteristics	**ES** (%)	**MS** (%)	**HS** (%)
Dictionary word	4.29	1.25	2.56
Dictionary word plus*	8.85	17.76	15.81
Numbers only*	31.64	13.08	8.12
(All other passwords)	55.22	67.91	73.51

Table 8. Passwords containing dictionary words or numbers

As in Table 8, only a small percentage (under 5 %) of all age groups) created passwords with a single dictionary word. There were significantly fewer ES created passwords using a single dictionary word plus numbers and special characters preceding or following the word– *Dictionary word plus*, as compared to their older counterparts–MS ($\chi^2 = 12.13$, df = 1) and HS ($\chi^2 = 10.19$, df = 1). There were significantly more ES (almost 1/3) created passwords with only numbers, as compared to MS ($\chi^2 = 33.47$, df = 1) and to HS ($\chi^2 = 98.83$, df = 1). In addition, significantly more MS created numbers-only passwords as compared to HS ($\chi^2 = 6.21$, df = 1). This indicates that as children progress from ES to HS, they created fewer and fewer numbers-only passwords.

The created passwords often consist of concepts reflecting the current state of the children's lives. Password themes included references to sports, video games, names, animals, movies, titles (princess, queen, etc.), numbers and colors. Passwords demonstrating these themes by ES include: "*12345*", "*Yellow*", "*doggysafesecure*", and "*PrincessFrog248*". Passwords created by MS include: "*Basketball1130*", "*GameGuy007*", and "*Gamehead77*". Passwords created by HS include: "*callofdutyblackops*", "*ILoveFortnite*", and "*Soccer player.15*". Several children provided their password creation strategies, instead of actually creating an example password. For instance, an ES wrote "*Maybe a birthdate or something.*" (P1168, 4th), another MS wrote "*My gamer tag, then random numbers*" (P2970, 8th), and an HS provided "*firstnamelastname123*" (P2837, 11th).

Password Strength

For the purpose of our study, we measured password strength with the password strength meter which uses the zxcvbn.js[4] script. This is an open-source tool, which uses pattern matching and searches for the minimum entropy of a given password. While we investigated the use of other password strength assessment tools, we were limited to tools that do not retain password data in order to comply with our IRB requirements.

The rating score provided by zxcvbn.js measures password strength on an ordinal scale with "0" being assigned to a password that can be guessed within 100 guesses. A "4" is assigned to a password that required over 10 to the power of 8 guesses. Collapsing password strength to a 5-item ordinal scale undeniably suppresses data variance. For example, if the number of guesses to crack one password was 1,100 and the estimated number of guesses for another password is 9,900, both passwords would be assigned a rating of 2. Yet there is a large difference in the number of guesses and the identical rating does not reflect this. Figure 7 shows the strengths of passwords across the three groups.

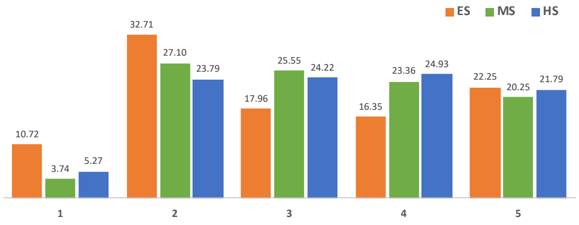


Figure 7. Password Strengths (in %)

The HS' passwords were significantly stronger than the ES' (z = 3.40). The MS' passwords were also significantly stronger (z = 2.42) than the ES'. For those passwords with a score of 1, the students used all numbers or simple common words as proposed passwords such as: "*1206*", "*112233*", "*Yellow*" and "*Game1234*". Examples of strong passwords (those with a score of 5) were:

- by ES: "*Love_Butter56*" and "*Dolphins blue tale*";
- by MS: "*ArrowTurner_8435!*" and "*dancingdinosaursavrwhoop164*";
- by HS: "*Soccer player.15*" and "*Aiken_bacon@28*".

5 Discussion

Not surprisingly, as children age, their use of technology and online activities change. The percentages of students having

[4] https://www.bennish.net/password-strength-checker/

cell phones increased almost 20 % from ES to MS and another 10 % from MS to HS. With age, social activities naturally increase as described in the PEW article of Teen, Social Media and Technology Study 2018 [2]. Our data confirm this trend—both texting and social media use increase significantly from ES to MS to HS. HS also use email significantly more than ES or MS. The increased technology use translates to needs for authentication for older children. A coping strategy may be that over 80 % of HS and MS reported using the same password for everything much like password reuse of adults [37, 42].

5.1 RQ1: Password Understandings

Generation Z, or those born from the mid-1990's to the late 2000's (the population of focus in this study) have several unique generational characteristics that influence their behavior [3] [29]. For example, they are digital natives and have grown up in a fully digital world where interaction with technologies is a part of normal life, requires authentication, and frequently involves personal information [29]. Additionally, more children are gaining access to a variety of technologies earlier and more frequently than their older counterparts, all of which are reflected in our participants' password understandings.

Participants frequently specifically mentioned securing their personal phones and computers, and were particularly concerned about access: the code *access* was applied to 601 participant responses, and pertained to both personal access to one's own devices/information and preventing unwanted access by others as seen in Figures 8, 9, and 10. For example, (P1880, 6th) indicated that one "*should have a password so that people won't go through your phone*" and (P394, 4th) found passwords to be important "*to unlock games (and) unlock computers*."

Frequently, *access* was associated with matters of *privacy*, as indicated in Figures 9 and 10 which demonstrate that MS and HS participants noted privacy concerns as their primary response. Whereas adults frequently worry about hackers' access to tangible things like bank account information, students frequently use technology for purposes deeply related to their identities like social media, gaming identities, and texting, and their password understandings reflect these uses. In terms of social development, as children–particularly preteens and teenagers like the majority of this study's participants–begin to explore and exercise autonomy, their privacy becomes an increasing concern. In this study, participants frequently emphasized the importance of passwords for personal information privacy, like (P2034, 11th) who commented that passwords "*secure...account(s) on social media*" and (P2972, 8th) who commented that passwords make it to where "*your siblings or family/friends can't get to any of your stuff*." Additionally, younger (ES) participants' privacy concerns were more general, whereas their MS and HS counterparts were increasingly more specific to things like gaming, social media, and cell phones. This makes sense, as younger students less frequently have unsupervised access to these applications and therefore do not associate them with expectations of privacy.

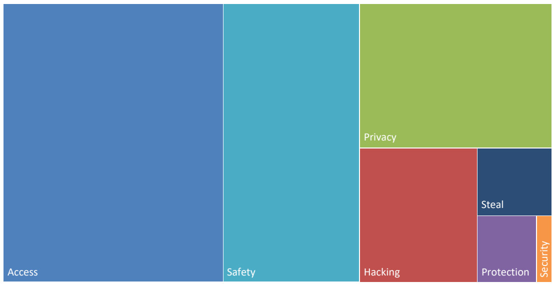


Figure 8. Why passwords? (ES)

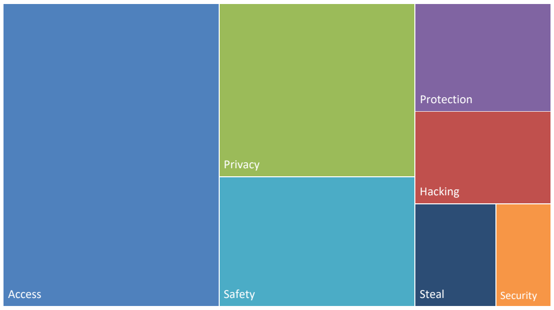


Figure 9. Why passwords? (MS)

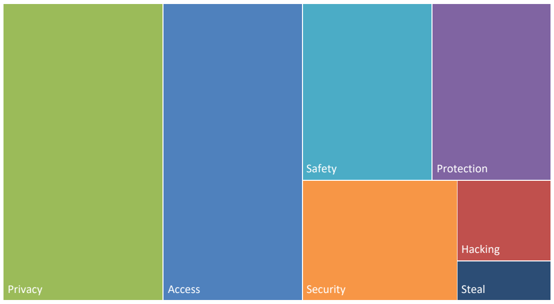


Figure 10. Why passwords? (HS)

Finally, though the idea of safety was an incredibly popular response in the open-ended question about students' password understandings (the words "safe" or "safety" appeared in 609 individual responses) the mentions of safety were, more than any other coded response, vague. For example, the words "safe" or "safety" were most likely to be written alone or accompanied by vague concepts like "things" and "stuff", e.g., "*to keep stuff safe*" (P1396, 11th) and "*to keep things safe*" (P1454, 7th). This raises questions about how much students really know about online/cybersecurity safety and privacy, and how much they have been raised in a digital age that teaches them that passwords and other security measures are important for safety, without ever explaining what that safety means. More open-ended qualitative investigation is needed to understand.

5.2 RQ2: Password Practices and Behaviors

Children's ages influence their password practices and behaviors. Younger children rely more on their family in creating and remembering passwords. Almost six times as many ES (about 90 %) reported having parental help in creating their passwords, in contrast to HS (about 15 %). Moreover, about 43 % of the younger children reported getting help from family members in remembering their passwords, as compared to only 7 % of the HS.

Both school and parents play an important role of providing guidance on 'good' password hygiene across all age groups. Additionally, almost half of MS and a third of HS reported assisting their family members with remembering passwords.

The participants reported having some good password behaviors including memorizing passwords, limiting writing passwords on paper, keeping their passwords private, and signing out after computer use (as shown in Figure 2 and Table 6). However, students in our study frequently used words (presumably) containing personal information, which is a less secure behavior that is also reflected in other studies of children's password behavior [10, 35]. Additionally, as students grew older, they were increasingly more likely to share their password(s) with friends. In the age of modern technology where at least 84% of teenagers own cell phones [11], this actually makes sense: the use of various in-phone applications, video, and camera functions is ubiquitous and socially casual. Some students share their phone passwords with close friends or significant others in order to establish trust and make access to certain phone functions faster and easier. Unfortunately, this behavior often stands in direct contradiction to the students' own perceptions that sharing passwords is bad.

The simplistic nature of passwords is expected for younger students where literacy is improving as they age. This is especially true with younger students who are working on mastering their alphabet and numbers. Special character use was very scarce across all of the grades. This is evidenced by the fact that very few special characters appeared in the passwords created by the children in this study. The overall use of special characters by ES was less than 0.75 % except for white space which had a frequency of 3.00 %. The few special characters used were common punctuation marks such as comma (,), period (.), dash (-), and exclamation (!).

Despite the awareness shown when discussing the purposes of passwords, the passwords chosen by the children (particularly by the younger age group) were weak. There were improvements in the older groups (both MS and HS are significantly stronger than ES). The MS and HS passwords are equally distributed among scores 2, 3, 4, 5 (Figure 7). Unfortunately, adults also create passwords that are weak and easy to guess [4, 12, 18, 28, 40, 41]. Generally, adults find it difficult to choose passwords that are easy to remember and hard to guess [43] especially given the overwhelming number of passwords they must manage [8, 14]. We did not ask students to explain why they chose the numbers, letters, and characters in their fabricated passwords.

There is clearly a need to address how children, particularly in the younger age group, understand and use passwords in regard to understanding threats to passwords and valuing accounts [38]. Children should be guided in discussions about password strength requirements and why these requirements exist. Traditional password requirements would suggest that the complexity and strength required should increase as the child's ability develops. However, new password guidelines published by the National Institute of Standards and Technology (NIST) state that password complexity requirements do not ensure strong passwords; instead, longer passphrase-like passwords are encouraged [15]. It will be helpful to provide guidance to youth on how to evaluate what it is that is being protected, how strong a password is needed, and how to create an appropriate password.

In addition, given the high level of password reuse of HS, it is also important to teach students of the risks of reuse and emphasize that having unique passwords is a more secure approach.

6 Limitations

Our study has several limitations which may limit the generalizability of our findings. First, our sample was a convenience sample based on geography and personal connections with schools. Future studies may use alternative participant recruitment in an effort to minimize potential bias. Second, the hypothetical password creation task can be viewed as contrived. However, it still provides invaluable insight on children's character choices and composition patterns in passwords. The final limitation is the use of self-report data. The youth respondents may have rationalized their behaviors by providing socially desirable explanations. Due to the study format–survey with brief short response questions–we weren't able to ask follow-up questions or ask students to elaborate on their responses or password creation choices. Future studies could use mixed method techniques, such as including interviews, to probe deeper into youths' perceptions on online security and privacy.

7 Conclusion

This study finds that children are not yet plagued by the overwhelming number of passwords that adults must manage. Children on average reported having two passwords for school and two to four passwords for home, while adults report having up to five times that amount [8, 14].

Reinforcing positive perceptions and practices

It is important to promote positive user perceptions about passwords early on [8], and our data indicates that children have reasonably accurate perceptions and knowledge of passwords and authentication. Thus, cybersecurity education should strive to reinforce these positive perceptions while continuing provide and promote security understandings.

Promoting concrete understanding

Our study also reveals that students frequently discuss the significance of passwords very generally and vaguely, often using one or two words like "information" and "safe," and do not put their password knowledge into practice. This raises questions about whether or not students actually understand why certain password practices exist versus just knowing about the practices. This, in turn, raises questions about whether or not, without this understanding, they will consistently make appropriate password choices across technologies and technological applications.

Bridging gap between knowledge and behavior

Further, this study reveals that children have appropriate perceptions and knowledge of passwords, but also demonstrate bad password habits that are contradictory to this knowledge. Students as young as third grade understand that passwords provide access controls, protect their privacy, and ensure their *stuff*'s safety. They also practice some good password practices such as memorizing passwords, limiting writing passwords down, keeping their passwords private, and logging out after sessions. However, many students exhibit password behaviors that do not align with their stated understanding of passwords, such as sharing passwords with friends, reusing passwords and using personal information when creating passwords.

This gap between students' stated password knowledge and their password behavior is an important next step for research surrounding children's password use and education. More mixed methods studies with more extensive questioning methods like interviews are needed to help better understand the nuances of children's perceptions of passwords, as well as the gap between knowledge and use. Understanding these nuances is important for thinking about how to better educate students about password behavior and online privacy and security, and how to move their knowledge into appropriate practice.

References

[1] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Communications of the ACM*, 42(12), 41-46.

[2] Monica Anderson, and Jingjing Jiang. Teens, social media & technology 2018. *Pew Research Center* 31 (2018): 2018. Retrieved September 17, 2019 from https://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/

[3] Sezin Baysal Berkup. 2014. Working with generations X and Y in generation Z period: Management of different generations in business life. *Mediterranean Journal of Social Sciences* 5.19 (2014): 218.

[4] Joseph Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *2012 IEEE Symposium on Security and Privacy.* 538–552. DOI:http://dx.doi.org/10.1109/SP.2012.49

[5] Charles P. Bourne and Donald F. Ford. 1961. A Study of the Statistics of Letters in English Words. *Information and Control*, 4(1): 48-67, 1961.

[6] Yee-Yin Choong. 2014. A cognitive-behavioral framework of user password management lifecycle. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 127–137. Springer, 2014.

[7] Yee-Yin Choong, Mary F. Theofanos, and Hung-Kung Liu. 2014. *United States Federal Employees' Password Management Behaviors: A Department of Commerce Case Study*. NISTIR 7991, 2014.

[8] Yee-Yin Choong and Mary F. Theofanos. 2015. What 4,500+ people can tell you–employees' attitudes toward organizational password policy do matter. In *International Conference on Human Aspects of Information Security, Privacy, and Trust,* pp. 299-310. Springer, Cham. 2015.

[9] Yee-Yin Choong, Mary F. Theofanos, Karen Renaud, and Suzanne Prior. 2019. Case Study–Exploring Children's Password Knowledge and Practices. In *Workshop on Usable Security and Privacy (USEC) 2019.*

[10] Porter E. Coggins III. 2013. Implications of what children know about computer passwords. *Computers in the Schools*, 30(3):282–293, 2013.

[11] Common Sense Media. 2019. The Common Sense Census: Media Use by Tweens and Teens, 2019. Retrieved from

https://www.commonsensemedia.org/Media-use-by-tweens-and-teens-2019-infographic

[12] Matteo Dell'Amico, Pietro Michiardi, and Yves Roudier. 2010. Password Strength: An Empirical Analysis. In *Proceedings of IEEE International Conference on Computer Communications (INFOCOM).*

[13] EU Kids Online. 2014. *EU Kids Online–Findings, methods, recommendations*. LSE, London: EU Kids Online. Available on http://lsedesignunit.com/EUKidsOnline.

[14] Dinei Florêncio and Cormac Herley. 2007. A Large-Scale Study of Web Password Habits. In: *Proceedings of the 16th International Conference on World Wide Web*, pp. 657-666. ACM, 2007.

[15] Paul Grassi, James L. Fenton, Elaine M. Newton, Ray A. Periner, Andrew R. Regensheid, William E. Burr, Justin P. Richer, Naomi B. Lefkovitz, Jamie M. Danker, Yee-Yin Choong, Kristen K. Greene, and Mary F. Theofanos. 2017. *Digital identity guidelines: Authentication and lifecycle management.* Technical Report 800-63B, NIST Special Publication, 2017.

[16] Libby Hanna, Kirsten Risden, and Kristin J. Alexander. 1997. Guidelines for usability testing with children. *interactions*, 4(5):9–14, 1997.

[17] Eiji Hayashi and Jason Hong. 2011. A Diary Study of Password Usage in Daily Life. In *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11*). ACM, New York, NY, USA, 2627–2630. DOI:http://dx.doi.org/10.1145/1978942.1979326.

[18] Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lo´pez. 2012. Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*. 523–537.

[19] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp.65-70. 1979.

[20] Gunther Kress. 1997. *Before writing: Rethinking the pathway into writing*. Routledge.

[21] Dev Raj Lamichhane and J C. Read. 2017. Investigating children's passwords using a game-based survey. In *Proceedings of the 2017 Conference on Interaction Design and Children*, IDC '17, pages 617–622, New York, NY, USA, 2017.

[22] Paul Y. Lee and Yee-Yin Choong. 2015. Human generated passwords–the impacts of password requirements and presentation styles. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 83–94. Springer, 2015.

[23] Michael P. Irwin. 2012. *An Investigation of Online Threat Awareness and Behaviour Patterns Amongst Secondary School Learners.* Doctoral dissertation, Rhodes University, Grahamstown, South Africa.

[24] Priya Kumar, Shalmali M. Naik, Utkarsha R. Devkar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. 2017. 'No Telling Passcodes Out Because They're Private': Understanding Children's Mental Models of Privacy and Security Online. *Proceedings of the ACM on Human-Computer Interaction*, *1,* CSCW, 64.

[25] Walter Loban. 1963. *The language of elementary school children.* National Council of Teachers of English, Champaign, IL, 1963.

[26] Stuart MacFarlane, Janet Read, Johanna Höysniemi, and Panos Markopoulos. 2003. Half-day tutorial: Evaluating interactive products for and with children. In *Interact*, pages 1027–1028, 2003.

[27] Sumbal Maqsood, Robert Biddle, Sana Maqsood, and Sonia Chiasson. 2018. An exploratory study of children's online password behaviours. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (pp. 539-544). ACM.

[28] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie F. Cranor, Patrick G. Kelley, Richard Shay, and Blase Ur. 2013. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security,* pp. 173-186. ACM, 2013.

[29] Oxford Royale Academy. 7 Unique Characteristics of Generation Z. (January 25, 2018). Retrieved September 06, 2019 from https://www.oxford-royale.co.uk/articles/7-unique-characteristics-generation-z.html

[30] Ugnė Paluckaitė, and Kristina Žardeckaitė-Matulaitienė. 2017. Adolescents' Perception of Risky Behaviour on the Internet. In *ICH&HPSY 2017: The European proceedings of social & behavioural sciences EpSBS: 3rd icH&Hpsy international conference on health and health psychology, July 5-7, 2017, Porto. London: Future Academy, 2017, vol. 30.*

[31] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. 2017. Let's

go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17*), 2017.

[32] Marc Prensky. 2001. Digital natives, digital immigrants. *On the Horizon,* 9(5), 2001. Retrieved from https://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf

[33] Janet C. Read, and Brendan Cassidy. 2012. Designing textual password systems for children. In *Proceedings of the 11th International Conference on Interaction Design and Children* (pp. 200-203). ACM.

[34] Karen Renaud and Joseph Maguire. 2015. Regulating access to adult content (with privacy preservation). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4019–4028. ACM, 2015.

[35] KwangCheol Rim, and SoYoung Choi. 2015. Analysis of Password Generation Types in Teenagers–Focusing on the Students of Jeollanam-do. *International Journal of u-and e-Service, Science and Technology*, *8*(9), 371-380.

[36] Johnny Saldaña. (2015) *The coding manual for qualitative researchers* (3rd Ed.). SAGE Publications.

[37] Elizabeth Stobert and Robert Biddle. 2014. The password life cycle: User behaviour in managing passwords. In *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS'14*), July 2014.

[38] The Digital Future Report. 2018. *The 16th annual study on the impact of digital technology on Americans***.** Center for the Digital Future at USC Annenberg , Retrieved September 17, 2019 from https://www.digitalcenter.org/wp-content/uploads/2018/12/2018-Digital-Future-Report.pdf

[39] United States Department of Education. 2019. *The condition of education, 2019.* National Center for Education Statistics. Retrieved September 21, 2020 from https://nces.ed.gov/pubs2019/2019144.pdf

[40] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2015. "I Added '!' at the End to Make It Secure": Observing password creation in the lab. In *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS'15*), 2015.

[41] Blase Ur, Patrick G. Kelley, Saranga Komanduri, Joel Lee, Michael Maass, Michelle L. Mazurek, Timothy Passaro, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. 2012. How does your password measure up? the effect of strength meters on password creation. In *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12),* pp. 65-80, 2012.

[42] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding password choices: How frequently entered passwords are re-used across websites. In *Proceedings of the 12th USENIX Conference on Us- able Privacy and Security (SOUPS '16*), 2016.

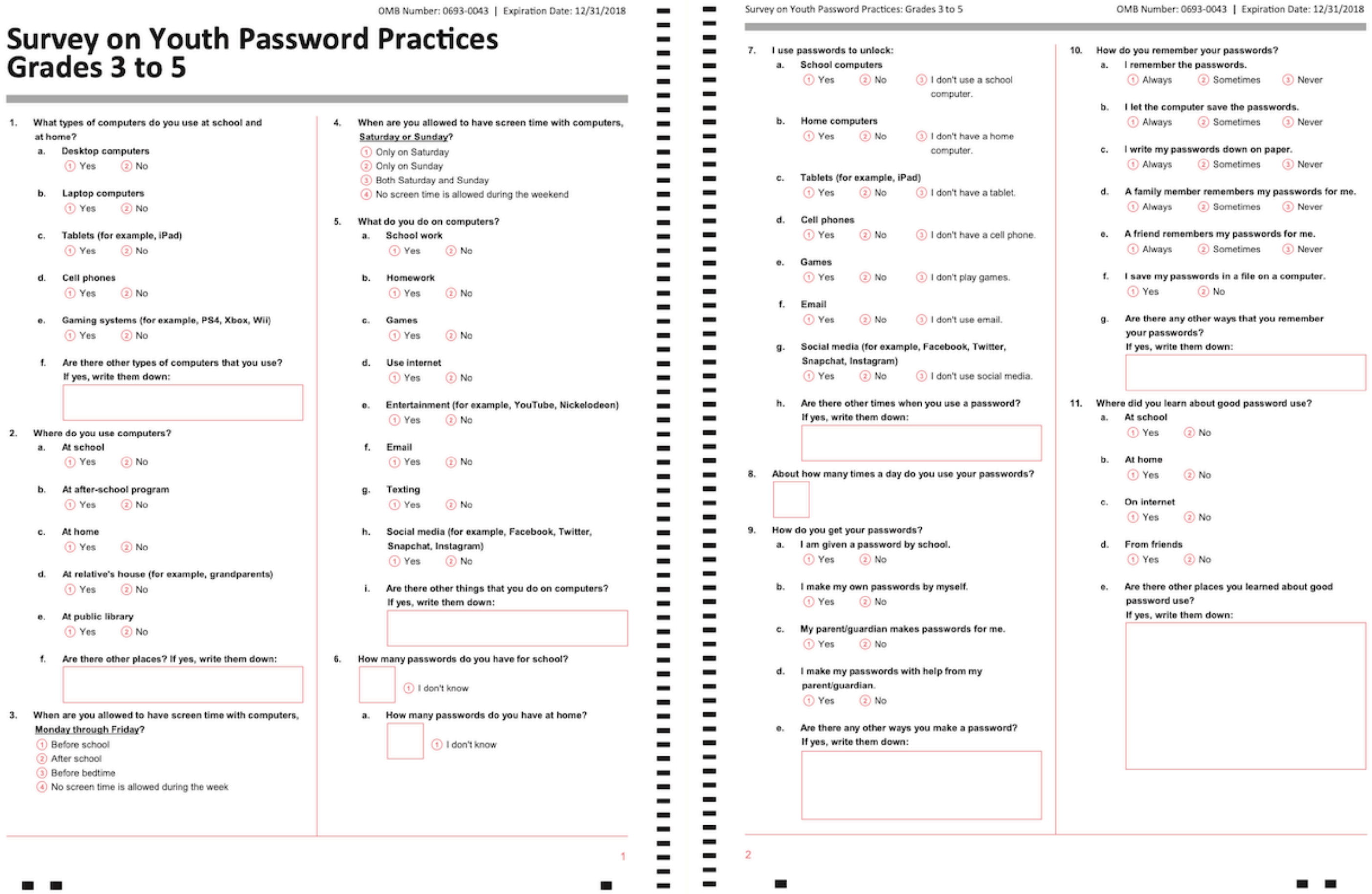
[43] Jeff Yan, Alan Blackwell, Ross Anderson, and Alasdair Grant. 2004. Password memorability and security: Empirical results. *IEEE Security and Privacy*, 2(5):25–31, September 2004.

[44] Ramazan Yilmaz, Fatma Gizem Karaoğlan Yilmaz, H. Tuğba Öztürk, and Tuğra Karademir. 2017. Examining Secondary School Students' Safe Computer and Internet Usage Awareness: an Example from Bartin Province. *Pegem Eğitim ve Öğretim Dergisi, 7*(1), 83-114.

[45] Leah Zhang-Kennedy, Christine Mekhail, Yomna Abdelaziz, and Sonia Chiasson. 2016. From Nosy Little Brothers to Stranger-Danger: Children and Parents' Perception of Mobile Threats. In *Proceedings of the The 15th International Conference on Interaction Design and Children* (pp. 388-399). ACM.

[46] Janko Žufić, Tomislava Žajgar, and S. Prkić. 2017. Children online safety. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 961-966). IEEE.

Appendix: Survey Instrument



OMB Number: 0693-0043 | Expiration Date: 12/31/2018

Survey on Youth Password Practices Grades 3 to 5

1. What types of computers do you use at school and at home?
   a. Desktop computers ① Yes ② No
   b. Laptop computers ① Yes ② No
   c. Tablets (for example, iPad) ① Yes ② No
   d. Cell phones ① Yes ② No
   e. Gaming systems (for example, PS4, Xbox, Wii) ① Yes ② No
   f. Are there other types of computers that you use? If yes, write them down: □

2. Where do you use computers?
   a. At school ① Yes ② No
   b. At after-school program ① Yes ② No
   c. At home ① Yes ② No
   d. At relative's house (for example, grandparents) ① Yes ② No
   e. At public library ① Yes ② No
   f. Are there other places? If yes, write them down: □

3. When are you allowed to have screen time with computers, Monday through Friday?
   ① Before school
   ② After school
   ③ Before bedtime
   ④ No screen time is allowed during the week

4. When are you allowed to have screen time with computers, Saturday or Sunday?
   ① Only on Saturday
   ② Only on Sunday
   ③ Both Saturday and Sunday
   ④ No screen time is allowed during the weekend

5. What do you do on computers?
   a. School work ① Yes ② No
   b. Homework ① Yes ② No
   c. Games ① Yes ② No
   d. Use internet ① Yes ② No
   e. Entertainment (for example, YouTube, Nickelodeon) ① Yes ② No
   f. Email ① Yes ② No
   g. Texting ① Yes ② No
   h. Social media (for example, Facebook, Twitter, Snapchat, Instagram) ① Yes ② No
   i. Are there other things that you do on computers? If yes, write them down: □

6. How many passwords do you have for school? □ ① I don't know
   a. How many passwords do you have at home? □ ① I don't know

1

Survey on Youth Password Practices: Grades 3 to 5

OMB Number: 0693-0043 | Expiration Date: 12/31/2018

7. I use passwords to unlock:
   a. School computers ① Yes ② No ③ I don't use a school computer.
   b. Home computers ① Yes ② No ③ I don't have a home computer.
   c. Tablets (for example, iPad) ① Yes ② No ③ I don't have a tablet.
   d. Cell phones ① Yes ② No ③ I don't have a cell phone.
   e. Games ① Yes ② No ③ I don't play games.
   f. Email ① Yes ② No ③ I don't use email.
   g. Social media (for example, Facebook, Twitter, Snapchat, Instagram) ① Yes ② No ③ I don't use social media.
   h. Are there other times when you use a password? If yes, write them down: □

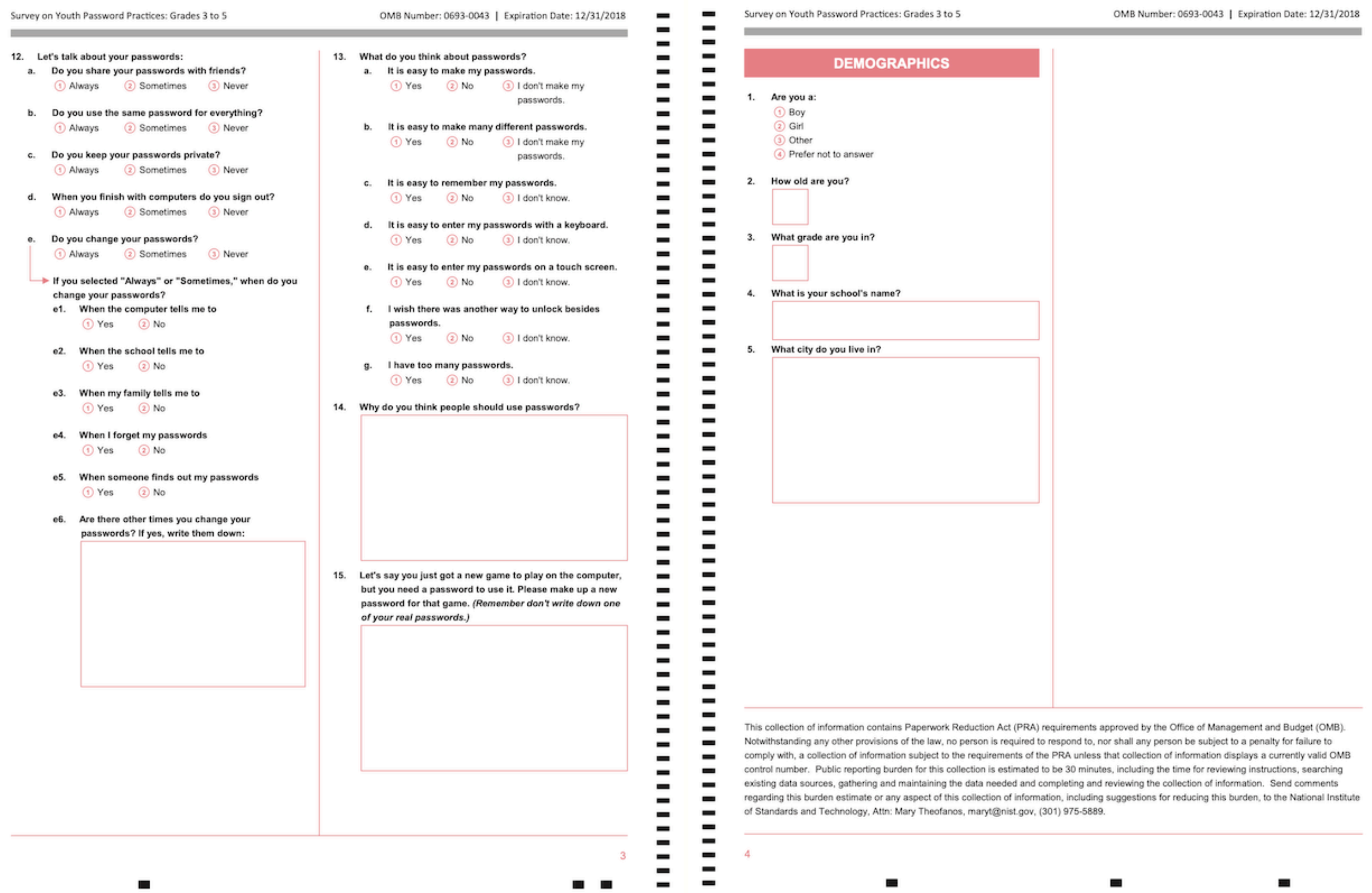
8. About how many times a day do you use your passwords? □

9. How do you get your passwords?
   a. I am given a password by school. ① Yes ② No
   b. I make my own passwords by myself. ① Yes ② No
   c. My parent/guardian makes passwords for me. ① Yes ② No
   d. I make my passwords with help from my parent/guardian. ① Yes ② No
   e. Are there any other ways you make a password? If yes, write them down: □

10. How do you remember your passwords?
   a. I remember the passwords. ① Always ② Sometimes ③ Never
   b. I let the computer save the passwords. ① Always ② Sometimes ③ Never
   c. I write my passwords down on paper. ① Always ② Sometimes ③ Never
   d. A family member remembers my passwords for me. ① Always ② Sometimes ③ Never
   e. A friend remembers my passwords for me. ① Always ② Sometimes ③ Never
   f. I save my passwords in a file on a computer. ① Yes ② No
   g. Are there any other ways that you remember your passwords? If yes, write them down: □

11. Where did you learn about good password use?
   a. At school ① Yes ② No
   b. At home ① Yes ② No
   c. On internet ① Yes ② No
   d. From friends ① Yes ② No
   e. Are there other places you learned about good password use? If yes, write them down: □

2



Survey on Youth Password Practices: Grades 3 to 5

OMB Number: 0693-0043 | Expiration Date: 12/31/2018

12. Let's talk about your passwords:
  a. Do you share your passwords with friends?
    ① Always ② Sometimes ③ Never
  b. Do you use the same password for everything?
    ① Always ② Sometimes ③ Never
  c. Do you keep your passwords private?
    ① Always ② Sometimes ③ Never
  d. When you finish with computers do you sign out?
    ① Always ② Sometimes ③ Never
  e. Do you change your passwords?
    ① Always ② Sometimes ③ Never

  If you selected "Always" or "Sometimes," when do you change your passwords?
    e1. When the computer tells me to
      ① Yes ② No
    e2. When the school tells me to
      ① Yes ② No
    e3. When my family tells me to
      ① Yes ② No
    e4. When I forget my passwords
      ① Yes ② No
    e5. When someone finds out my passwords
      ① Yes ② No
    e6. Are there other times you change your passwords? If yes, write them down:

13. What do you think about passwords?
  a. It is easy to make my passwords.
    ① Yes ② No ③ I don't make my passwords.
  b. It is easy to make many different passwords.
    ① Yes ② No ③ I don't make my passwords.
  c. It is easy to remember my passwords.
    ① Yes ② No ③ I don't know.
  d. It is easy to enter my passwords with a keyboard.
    ① Yes ② No ③ I don't know.
  e. It is easy to enter my passwords on a touch screen.
    ① Yes ② No ③ I don't know.
  f. I wish there was another way to unlock besides passwords.
    ① Yes ② No ③ I don't know.
  g. I have too many passwords.
    ① Yes ② No ③ I don't know.

14. Why do you think people should use passwords?

15. Let's say you just got a new game to play on the computer, but you need a password to use it. Please make up a new password for that game. *(Remember don't write down one of your real passwords.)*

3

Survey on Youth Password Practices: Grades 3 to 5

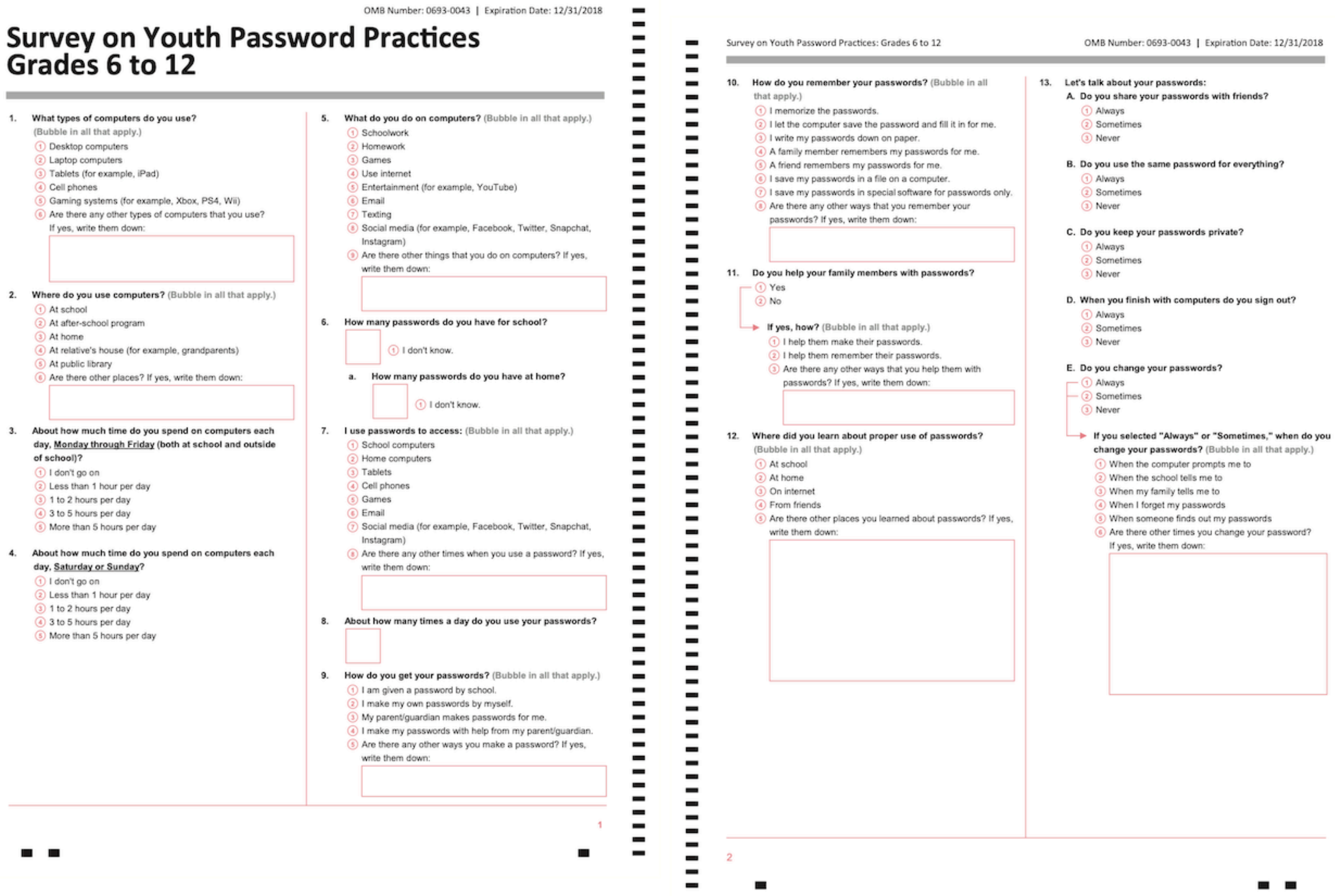
OMB Number: 0693-0043 | Expiration Date: 12/31/2018

DEMOGRAPHICS

1. Are you a:
  ① Boy
  ② Girl
  ③ Other
  ④ Prefer not to answer
2. How old are you?
3. What grade are you in?
4. What is your school's name?
5. What city do you live in?

This collection of information contains Paperwork Reduction Act (PRA) requirements approved by the Office of Management and Budget (OMB). Notwithstanding any other provisions of the law, no person is required to respond to, nor shall any person be subject to a penalty for failure to comply with, a collection of information subject to the requirements of the PRA unless that collection of information displays a currently valid OMB control number. Public reporting burden for this collection is estimated to be 30 minutes, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed and completing and reviewing the collection of information. Send comments regarding this burden estimate or any aspect of this collection of information, including suggestions for reducing this burden, to the National Institute of Standards and Technology, Attn: Mary Theofanos, maryt@nist.gov, (301) 975-5889.

4



OMB Number: 0693-0043 | Expiration Date: 12/31/2018

Survey on Youth Password Practices Grades 6 to 12

1. **What types of computers do you use?** (Bubble in all that apply.)
   ① Desktop computers
   ② Laptop computers
   ③ Tablets (for example, iPad)
   ④ Cell phones
   ⑤ Gaming systems (for example, Xbox, PS4, Wii)
   ⑥ Are there any other types of computers that you use? If yes, write them down:

2. **Where do you use computers?** (Bubble in all that apply.)
   ① At school
   ② At after-school program
   ③ At home
   ④ At relative's house (for example, grandparents)
   ⑤ At public library
   ⑥ Are there other places? If yes, write them down:

3. **About how much time do you spend on computers each day, Monday through Friday (both at school and outside of school)?**
   ① I don't go on
   ② Less than 1 hour per day
   ③ 1 to 2 hours per day
   ④ 3 to 5 hours per day
   ⑤ More than 5 hours per day

4. **About how much time do you spend on computers each day, Saturday or Sunday?**
   ① I don't go on
   ② Less than 1 hour per day
   ③ 1 to 2 hours per day
   ④ 3 to 5 hours per day
   ⑤ More than 5 hours per day

5. **What do you do on computers?** (Bubble in all that apply.)
   ① Schoolwork
   ② Homework
   ③ Games
   ④ Use internet
   ⑤ Entertainment (for example, YouTube)
   ⑥ Email
   ⑦ Texting
   ⑧ Social media (for example, Facebook, Twitter, Snapchat, Instagram)
   ⑨ Are there other things that you do on computers? If yes, write them down:

6. **How many passwords do you have for school?**
   ① I don't know.

   a. **How many passwords do you have at home?**
   ① I don't know.

7. **I use passwords to access:** (Bubble in all that apply.)
   ① School computers
   ② Home computers
   ③ Tablets
   ④ Cell phones
   ⑤ Games
   ⑥ Email
   ⑦ Social media (for example, Facebook, Twitter, Snapchat, Instagram)
   ⑧ Are there any other times when you use a password? If yes, write them down:

8. **About how many times a day do you use your passwords?**

9. **How do you get your passwords?** (Bubble in all that apply.)
   ① I am given a password by school.
   ② I make my own passwords by myself.
   ③ My parent/guardian makes passwords for me.
   ④ I make my passwords with help from my parent/guardian.
   ⑤ Are there any other ways you make a password? If yes, write them down:

1

Survey on Youth Password Practices: Grades 6 to 12 OMB Number: 0693-0043 | Expiration Date: 12/31/2018

10. **How do you remember your passwords?** (Bubble in all that apply.)
    ① I memorize the passwords.
    ② I let the computer save the password and fill it in for me.
    ③ I write my passwords down on paper.
    ④ A family member remembers my passwords for me.
    ⑤ A friend remembers my passwords for me.
    ⑥ I save my passwords in a file on a computer.
    ⑦ I save my passwords in special software for passwords only.
    ⑧ Are there any other ways that you remember your passwords? If yes, write them down:

11. **Do you help your family members with passwords?**
    ① Yes
    ② No

    If yes, how? (Bubble in all that apply.)
    ① I help them make their passwords.
    ② I help them remember their passwords.
    ③ Are there any other ways that you help them with passwords? If yes, write them down:

12. **Where did you learn about proper use of passwords?** (Bubble in all that apply.)
    ① At school
    ② At home
    ③ On internet
    ④ From friends
    ⑤ Are there other places you learned about passwords? If yes, write them down:

13. **Let's talk about your passwords:**

    A. **Do you share your passwords with friends?**
    ① Always
    ② Sometimes
    ③ Never

    B. **Do you use the same password for everything?**
    ① Always
    ② Sometimes
    ③ Never

    C. **Do you keep your passwords private?**
    ① Always
    ② Sometimes
    ③ Never

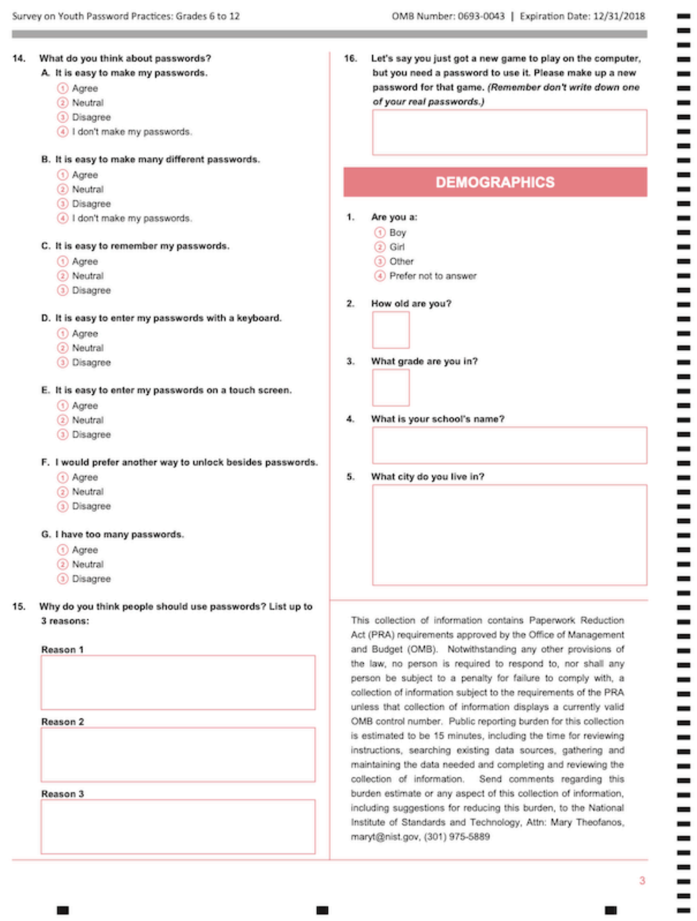
    D. **When you finish with computers do you sign out?**
    ① Always
    ② Sometimes
    ③ Never

    E. **Do you change your passwords?**
    ① Always
    ② Sometimes
    ③ Never

    If you selected "Always" or "Sometimes," when do you change your passwords? (Bubble in all that apply.)
    ① When the computer prompts me to
    ② When the school tells me to
    ③ When my family tells me to
    ④ When I forget my passwords
    ⑤ When someone finds out my passwords
    ⑥ Are there other times you change your password? If yes, write them down:

2

Theofanos, Mary; Choong, Yee-Yin. "'Passwords Keep Me Safe' – Understanding What Children Think about Passwords." Presented at 30th USENIX Security Symposium. August 11, 2021 - August 13, 2021.



Survey on Youth Password Practices: Grades 6 to 12

OMB Number: 0693-0043 | Expiration Date: 12/31/2018

14. What do you think about passwords?

A. It is easy to make my passwords.
- ① Agree
- ② Neutral
- ③ Disagree
- ④ I don't make my passwords.

B. It is easy to make many different passwords.
- ① Agree
- ② Neutral
- ③ Disagree
- ④ I don't make my passwords.

C. It is easy to remember my passwords.
- ① Agree
- ② Neutral
- ③ Disagree

D. It is easy to enter my passwords with a keyboard.
- ① Agree
- ② Neutral
- ③ Disagree

E. It is easy to enter my passwords on a touch screen.
- ① Agree
- ② Neutral
- ③ Disagree

F. I would prefer another way to unlock besides passwords.
- ① Agree
- ② Neutral
- ③ Disagree

G. I have too many passwords.
- ① Agree
- ② Neutral
- ③ Disagree

15. Why do you think people should use passwords? List up to 3 reasons:

Reason 1

Reason 2

Reason 3

16. Let's say you just got a new game to play on the computer, but you need a password to use it. Please make up a new password for that game. *(Remember don't write down one of your real passwords.)*

DEMOGRAPHICS

1. Are you a:
- ① Boy
- ② Girl
- ③ Other
- ④ Prefer not to answer

2. How old are you?

3. What grade are you in?

4. What is your school's name?

5. What city do you live in?

This collection of information contains Paperwork Reduction Act (PRA) requirements approved by the Office of Management and Budget (OMB). Notwithstanding any other provisions of the law, no person is required to respond to, nor shall any person be subject to a penalty for failure to comply with, a collection of information subject to the requirements of the PRA unless that collection of information displays a currently valid OMB control number. Public reporting burden for this collection is estimated to be 15 minutes, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed and completing and reviewing the collection of information. Send comments regarding this burden estimate or any aspect of this collection of information, including suggestions for reducing this burden, to the National Institute of Standards and Technology, Attn: Mary Theofanos, maryt@nist.gov, (301) 975-5889

3

Storing and retrieving wavefronts with resistive temporal memory

Advait Madhavan*, Mark D. Stiles†
Physical Measurement Laboratory, National Institute of Standards and Technology*†
Institute for Research in Electronics and Applied Physics, University of Maryland, College Park*
Email: *advait.madhavan@nist.gov, †mark.stiles@nist.gov

***Abstract*—We extend the reach of temporal computing schemes by developing a memory for multi-channel temporal patterns or "wavefronts." This temporal memory re-purposes conventional one-transistor-one-resistor (1T1R) memristor crossbars for use in an arrival-time coded, single-event-per-wire temporal computing environment. The memristor resistances and the associated circuit capacitances provide the necessary time constants, enabling the memory array to store and retrieve wavefronts. The retrieval operation of such a memory is naturally in the temporal domain and the resulting wavefronts can be used to trigger time-domain computations. While recording the wavefronts can be done using standard digital techniques, that approach has substantial translation costs between temporal and digital domains. To avoid these costs, we propose a spike timing dependent plasticity (STDP) inspired wavefront recording scheme to capture incoming wavefronts. We simulate these designs with experimentally validated memristor models and analyze the effects of memristor non-idealities on the operation of such a memory.**

I. The need for Temporal Memory

The three pillars that form the foundation of any computing system are computation (for processing), input/output (I/O) (for sensing and feedback), and memory (for storage). In the context of single-spike-per-wire, arrival-time coded computation, circuits that allow sensing and processing natively in the temporal domain are already being researched. Dynamic vision sensor (DVS) cameras [1], time-to-first-spike (TTFS) vision sensors [2], and address event representation (AER) ears [3] are a few examples of the sensing systems that natively encode information in the temporal domain. On the computational side, a space-time computing approach [4] has been proposed as a novel paradigm that encodes information in the relative arrival time between input events. A direct implementation of such a paradigm with off the shelf complementary-metal-oxide-semiconductor(CMOS) technology has also been proposed and demonstrated [5], [6]. The active research on two of the three pillars increases the urgency to develop of a memory that natively operates in the time domain. As of now, no such memory exists.

The biological motivation behind single-spike-per-wire temporal computation can be traced back to Thorpe and Imbert's work [7] arguing that the speed of processing of the visual system is too fast for a rate-coded interpretation of neural computation to be feasible [8]. Instead, they proposed a wavefront based computing approach that encodes information in the relative arrival time between a volley of spikes [9] as is shown in Fig. 1(a). This information representation is radically different from the conventional Boolean one and opens up a vastly different trade-off space of possible computing architectures [4]–[6], [10]–[12].

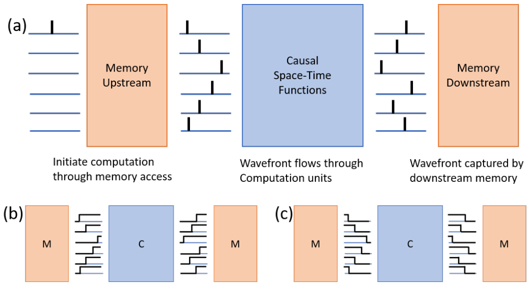

(a)
Memory Upstream
Causal Space-Time Functions
Memory Downstream
Initiate computation through memory access
Wavefront flows through Computation units
Wavefront captured by downstream memory
(b)
M
C
M
(c)
M
C
M

Fig. 1. A temporal computation procedure: Panel (a) shows an upstream memory through which computation is initiated by triggering an event at the input at a given time. The memory outputs a sequence of events on different lines which encodes the temporal wavefront that is stored in the first memory location. The output wavefront from the computational unit is read in by the downstream memory and stored in a memory location of choice. Panels (b) and (c) depict race-logic-like implementations that represent events with digital rising or falling edges instead of spikes.

The temporal domain allows for different encoding schemes, among which two have been well studied [13]–[15]. One is a timing code, where the exact delays between the spikes carry information [14], [16]. A more relaxed version of such an approach, though less dense in coding space, is a rank order code, in which only the relative orderings of the spikes carry information [15], [16]. Though sparser than their exact timing counterparts, order codes have been shown to contain appreciable information capacity while still maintaining robustness to variations in individual spike timing [15]. Though the first temporal networks were studied as early as 1990 [13], more recently, rank order codes have been studied in the context of deep spiking neural networks [17]–[19]. Such single-event-per-wire based (also known as non-leaky integrate and fire) models have been trained with modified STDP algorithms. Early results [18], [20] report comparable accuracies to deep learning networks but with small network sizes.

Building hardware implementations of such systems requires physical realizations of events. Race logic [5] is a temporally coded logic family that takes advantage of the simplicity of the digital domain and represents events with

Madhavan, Advait; Stiles, Mark D. "Storing and retrieving wavefronts with resistive temporal memory." Presented at International Symposium on Circuits and Systems (ISCAS 2020). October 10, 2020 - October 21, 2020.

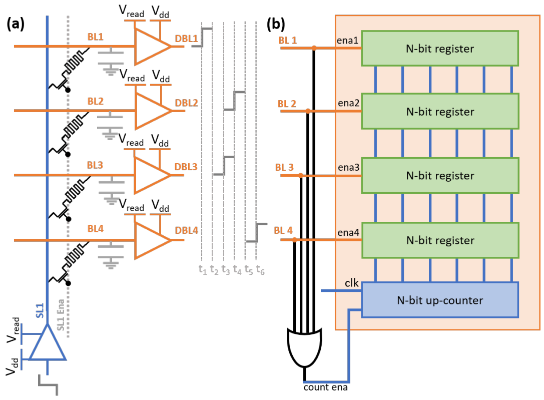


Fig. 2. A simple temporal memory circuit: Panel (a) shows a read/recall operation where a rising edge is presented at the input of the source line driver (level shifter). This results in different charging rates of the various bit lines, determined by the respective cross-point devices, resulting in an output wavefront. This wavefront enters the bitlines of panel (b) where the OR gate detects the first arriving edge, starting the up-counter. The incoming rising edges on the bit-lines latch the counted values and hence store the incoming wavefront.

rising or falling digital edges (as shown in Figs. 1(b,c)), instead of spikes. Fig. 1(b,c) show rising and falling edge versions of architectures of such a temporal computer, built using the race logic encoding scheme. Computations can be initiated by memory access to the upstream memory, which recalls a stored temporal wavefront. This wavefront flows through computational units that implement arbitrary causal functions such as the ones described in [4], [10], [11]. Lastly the downstream memory gets triggered with the first arriving edge and captures the incoming wavefront.

In this paper we present the design of a memory that fits seamlessly into a temporal computation procedure as described in figure 1. We do this by performing a translation between static memory and timing signals, through tunable memristor RC time constants. Section II describes how 1T1R memristive crossbars can be used to create wavefronts that have been stored in them. We describe how such an approach can interrogate the memristive state with more energy efficiency than conventional techniques. Section III describes how the relative timing information in a wavefront can be captured through standard digital circuit techniques which then invokes specific write circuitry to tune the memristor resistances to the corresponding captured digital values. This domain shifting, from analog to digital and back, has significant overhead associated with it. We then describe a proposed solution to natively capture wavefronts directly into memristors.

II. Recalling wavefronts stored in a memristor crossbar

An ideal temporal memory would be one that could be directly interfaced with digital components in a temporally coded environment where rising edges are used to demarcate events. Figure 2(a) shows a single column of such a memory, which uses a 1T1R memristor crossbar as its fundamental component. Each row behaves as an output bit line and each column behaves as the input source line. When a rising edge arrives through an enabled source line, it charges the output bit line (BL) MOS capacitor (shown in Fig. 2), through the memristor, until a threshold is reached, causing a rising edge at the digital bit line (DBL). Using such a circuit, the values of the memristive states can be directly read out at as a wavefront of digital rising edges, also known as wavefront recalling. This is shown in Fig. 2, where a linear variation in memristive values leads to a linearly spaced output wavefront.

Though the structure of the crossbar remains the same, the way it is used in this work differs in some important ways from conventional approaches. When used in a multilevel memory or feed-forward inference context, a static read voltage is applied across the device (pinned with a sense/measurement amplifier) while the resultant current is summed and measured. Hence, the energy efficiency in these approaches improves the larger the R_{on} and R_{off} resistances become. In contrast, in this RC charging based recall mode of operation, the voltage drop across the device is not static, because the voltage on the output capacitor changes during a read operation (Fig. 4(b)(iii)).

This changing voltage has two advantages. First, and more important, it decouples the energy cost per read operation from the value stored in the memristor. Independent of the state of the device, a single read operation consumes $CV_{\mathrm{read}}^2 (\approx 600\ \mathrm{fJ})$ of energy per line, with $CV_{\mathrm{read}}^2/2$ lost due to joule heating across the memristor and $CV_{\mathrm{read}}^2/2$ stored on the capacitor. This data independent energy cost allows memristors to be used in the high conductance regime, without incurring the increased energy cost. Circuit and architectural designs can then take advantage of the high conductance regime, where the behaviour of the device is more linear, repeatable and less susceptible to variation. Recently, for very low resistance states, the device to device variation has shown to be $\leq 1\%$ [21]. The second advantage is that the degree of read disturb on the device is reduced as the full read voltage is applied across the device for a relatively short period of time.

To enable easy interface with digital signal levels, level-shifters are required to translate between the memristor read voltages (V_{read}) and digital voltage levels (V_{dd}). This shifting down process can be implemented with regular inverters but the shifting up process requires either current mirror based level-shifters or cross coupled level-shifters. The current mirror based designs have a smoother response, and consume static power while the cross coupled versions are more power efficient, but have a more complicated response.

The cross coupled topology is representative of a positive feedback loop between transistors M1-M4 (Fig. 3). This positive feedback loop itself has a time constant that varies with the current charging the input node. This variable time constant can add timing uncertainties that are data dependent and could cause errors. One way to avoid this problem is to take advantage of the one sided nature of this information encoding. Using rising edges only determines the transistor that is responsible for the pull-down so it can be sized

Madhavan, Advait; Stiles, Mark D. "Storing and retrieving wavefronts with resistive temporal memory." Presented at International Symposium on Circuits and Systems (ISCAS 2020). October 10, 2020 - October 21, 2020.

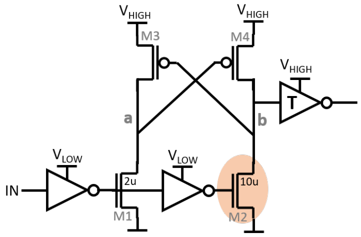


Fig. 3. Asymmetric rising edge level shifter: Here transistor M2 is sized larger than its counterpart M1 such that node "b" is pulled down faster with little competition from M1 via node "a". The inverter with a "T" inside represents a tri-state buffer.

accordingly larger. This approach makes the response of the level shifter more uniform.

III. Capturing Wavefronts: Digital vs Native

A functionally correct digital timing measurement approach to record wavefronts is shown in Fig. 2(b). High speed up-counters can be used for time scales on the order of 1 ns to 50 ns, while vernier delay lines, which extend the precision to the order of a single-inverter-delay, can be used for more precise measurements [22]. Using race logic principles, the first arriving edge is detected with an OR gate, which signals the beginning of the timing measurement system (counter or vernier delay line). With each subsequently arriving rising edge, the corresponding count value is captured in a register bank. An AND gate signals the last arriving input, at which time the recording process comes to an end with a digital representation of the relative arrival times with stored in a temporary register bank. These values can be used as targets for a closed loop feedback programming approach [23] that writes the corresponding values into the correct memory column.

To increase the energy efficiency of wavefront recording we eliminate the need to translate between encoding domains by using the ability to change memristor resistances with applied voltage pulses. This native approach to capturing wavefronts, results in a more natural and energy efficient, albeit more error prone implementation. In a time coded information representation, in which the plastic memristor resistances explicitly encode tunable delays, STDP-like behaviour can be used to record wavefronts as shown in Fig. 4. In this approach, the first arriving edge is conceptually treated as the "post" edge. The circuit then applies backward pulses of variable lengths across the memristors proportional to the *difference in timing* between this first-arriving "post" event and the later-arriving events, which can be thought of as "pre" events. The device with the largest difference between "pre" and "post" events, has the maximum conductance change, and hence the highest resistance. When a wavefront is then recalled, the highest conductance device responds first and the most resistive one responds last, preserving the wavefront shape.

Simulation results for such a procedure are shown in Fig. 4(b). These simulations are performed in a 180 nm process node, with a 1.8 V power supply. The memristor models used are from [24], and are modelled based on experimental measurements reported in [24], [25]. The wavefront recording operation proceeds by first initializing the column in question, (column 1, shown in figure 4(a)), with all memristors set to the ON state($\approx$ 10 kΩ) and the enable line (Ena1) activated. This can be seen in the first 100 ns of figure 4(b)(v) with all devices having the same impedance. The write path through the multiplexers, as shown in figure 4(a), is also activated, such that the OR gate controls the source line (SL).

The wavefront (having a dynamic range of 40 ns) to be recorded is presented at the digital bit lines (DBLs), which behave like the input in this phase of operation. Similarly to the digital case, the first arriving rising edge is detected by an OR gate, which triggers the application of an appropriate write voltage ($V_{\text{write}} \approx 1.4$ V), through the multiplexer, to the source line (SL). The bit-lines (BLs) of the array are operated in the write voltage regime with rising edges level shifted down from V_{dd} to V_{write}. Each device sees the difference in voltage between the source line(figure 4(b)(iv)) and corresponding bit lines(figure 4(b)(iii)) applied across it. For the device corresponding to the first arriving edge, both its source line and bit line go high at the same time, so there is no change in the memristive state. Meanwhile, the other devices experience a reverse V_{write} voltage across them, since their edges haven't arrived yet. They experience this voltage for the difference in time between their corresponding edge and the first arriving edge, hence causing a change in the memristive state proportional to the relative times between the inputs.

Once appropriate pulse lengths have been successfully applied across the devices, a copy of the input wavefront should be captured into the memristive device states. The last arriving edge signals the end of the recording operation and the circuit is reset with an external reset pulse. The reset pulse (not shown in the figures), discharges the crossbar without affecting the device state. The array is now ready for playback.

IV. Discussion

While such a wavefront recording approach seems feasible, some problems arise in the context of exact timing codes. First, the relationship between memristor resistance and relative output timings for recalling the wavefront is linear, arising directly from the $t \propto RC$ relationship. On the other hand, for recording the wavefront, the relationship between memristor conductance and voltage pulse duration is not linear, and depends on material properties. Since most memristive state change dynamics are governed by transitions over energy barriers, the effectiveness of a fixed voltage to change the device state drops logarithmically. In the wavefront recording process, a linearly spaced input wavefront will end up creating a logarithmically spaced resistive change, which when recalled would create a logarithmically spaced output wavefront. This problem is fundamental, being governed by the exponential nature of Boltzmann statistics and energy barriers.

In order to get linear behavior out of such a device, it must operate in a regime where the Taylor series expansion of its

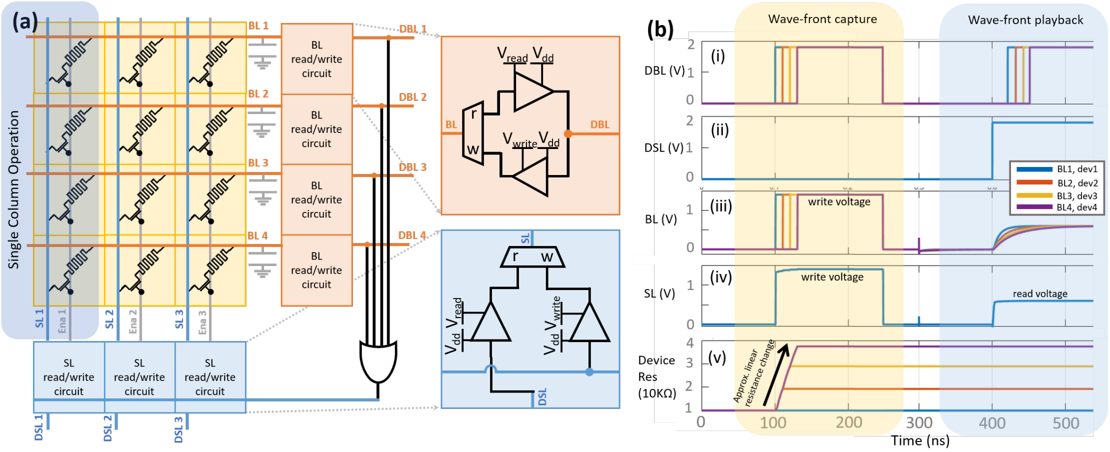


Fig. 4. Simulation results for recalling and recording wavefronts with resistive temporal memory: Panel (a) shows a 3X4 resistive temporal memory with the yellow cells representing the conventional 1T1R array, while the blue cells and orange cells represent the source and bit line augmentations that allow using such an array in a temporal context. Note that the level shifters shown in the zoomed in cells are digital, with the read out cells being tri-state, while the multiplexers that explicitly depict the read and write circuit paths are pass gate based. Panel (b) shows the SPICE simulation results of recording a wavefront and subsequently recalling it for the single column in Panel (a) that is highlighted in blue. The wavefronts are superimposed over each other to save space. In the capture phase, the bit-lines (BLs) are used as inputs, while the source lines(SLs) are controlled by the OR gate through the write path. In the recall phase, the source lines are used as the inputs with the bit lines being the outputs. Panels (i) and (ii) show the digital signal values, while panels (iii) and (iv) show the internal BL and SL values. Panel (v) shows the device state change during the capture phase.

behavior has small higher order coefficients, so that it can be approximated as linear. Such behavior can be seen for a range of voltages where a sharp pulse ($\leq$ 40 ns) across the device creates a linear change in the devices state (from 10 kΩ to 40 kΩ), which tapers off if the pulse is applied for a longer time. Here, as shown in figure 4(e)(v), our pulse duration is calibrated to access that specific linear region of the memristor IV characteristics, and therefore does not access the complete available dynamic range.

The reduced range is not detrimental and depends on the quality of the memristive devices being used. Multiple groups have shown 5 bit or more resolution in limited resistance ranges with the low resistance state incurring, programming cycle to cycle variations as low as $4.2\ \%$ and device to device variation as low as $4.5\ \%$ [26]. For very low resistances (between 1 kΩ and 10 kΩ), even lower variation numbers have been reported ($\leq 1\ \%$ [21]). Such technological improvements allow us to extract 4 to 5 bits of precision, even from a reduced dynamic range.

A second difficulty for exact timing codes is that the time scales of recording and of recalling need to match. For example, the resistance change created by 10 ns pulses in the recording process, should create 10 ns spaced edges when recalled. While the former is a material property and cannot be changed by circuit techniques, the latter can be addressed by adding a digitally programmable capacitances ($\approx$ 1 pF, in the current simulation) on the output line to correctly scale the timescale. For small array sizes such a capacitance can take up extra area, but as the array is scaled to more representative sizes, the crossbar, transistor-drain and driver capacitances will contribute significantly to this capacitance. Array scaling will also require scaling of the array drive circuits, especially with the high conductance regime operation. Though the write drivers on the bit line do not need to be adjusted, but the source line write driver will have to be designed to support N memristors in parallel during the capture phase. Future work includes a more detailed scaling analysis accounting for crossbar wire resistances and capacitances.

An important point to note is that rank order codes are more tolerant to the aforementioned concerns than exact timing codes. Logarithmic compression preserves order, and variable capacitances can be used with order codes to stretch the logarithmically compressed stored values. This allows enough write pulse duration to still change state on the next write operation. This makes rank order codes a more robust and error tolerant encoding for this kind of a temporal memory.

V. Conclusion

In this work we have proposed and validated through simulation, a single-event-per-wire temporal memory that operates in the sub 50 ns timing range while utilizing the low variability, low resistance states (10 kΩ to 40 kΩ) of memristive devices. We show how our recalling/playback operation has an energy cost of about 600 fJ per line, whose magnitude is independent of the device conductance. Rank order coded architectures seem to be the more promising encoding schemes for such an approach due to their error tolerance. Though many challenges remain, we believe that this is a first step towards realizing temporal memories that can work synergistically with tomorrow's temporally coded computing architectures.

References

[1] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128× 128 120 db 15μs latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] J. A. Lenero-Bardallo, T. Serrano-Gotarredona, and B. Linares-Barranco, "A signed spatial contrast event spike retina chip," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 2010, pp. 2438–2441.

[3] V. Chan, S.-C. Liu, and A. van Schaik, "Aer ear: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, 2007.

[4] J. E. Smith, "Space-time algebra: a model for neocortical computation," in *Proceedings of the 45th Annual International Symposium on Computer Architecture*. IEEE Press, 2018, pp. 289–300.

[5] A. Madhavan, T. Sherwood, and D. Strukov, "Race logic: A hardware acceleration for dynamic programming algorithms," in *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*. IEEE, 2014, pp. 517–528.

[6] A. Madhavan, T. Sherwood, and D.Strukov, "A 4-mm 2 180-nm-cmos 15-giga-cell-updates-per-second dna sequence alignment engine based on asynchronous race conditions," in *2017 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2017, pp. 1–4.

[7] S. J. Thorpe and M. Imbert, "Biological constraints on connectionist modelling," *Connectionism in perspective*, pp. 63–92, 1989.

[8] S. J. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *nature*, vol. 381, no. 6582, p. 520, 1996.

[9] S. Thorpe, A. Delorme, and R. Van Rullen, "Spike-based strategies for rapid processing," *Neural networks*, vol. 14, no. 6-7, pp. 715–725, 2001.

[10] G. Tzimpragos, A. Madhavan, D. Vasudevan, D. Strukov, and T. Sherwood, "Boosted race trees for low energy classification," in *Proceedings of the Twenty-Forth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '19, April 2019.

[11] M. H. Najafi, D. J. Lilja, M. D. Riedel, and K. Bazargan, "Low-cost sorting network circuits using unary processing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 8, pp. 1471–1480, 2018.

[12] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "Hots: a hierarchy of event-based time-surfaces for pattern recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1346–1359, 2016.

[13] S. J. Thorpe, "Spike arrival times: A highly efficient coding scheme for neural networks," *Parallel Processing in Neural Systems*, pp. 91–94, 1990.

[14] R. VanRullen, R. Guyonneau, and S. J. Thorpe, "Spike times make sense," *Trends in neurosciences*, vol. 28, no. 1, pp. 1–4, 2005.

[15] S. Thorpe and J. Gautrais, "Rank order coding," in *Computational neuroscience*. Springer, 1998, pp. 113–118.

[16] R. V. Rullen and S. J. Thorpe, "Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex," *Neural computation*, vol. 13, no. 6, pp. 1255–1283, 2001.

[17] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural Networks*, 2018.

[18] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier, "Stdp-based spiking deep convolutional neural networks for object recognition," *Neural Networks*, vol. 99, pp. 56–67, 2018.

[19] M. Mozafari, M. Ganjtabesh, A. Nowzari-Dalini, S. J. Thorpe, and T. Masquelier, "Bio-inspired digit recognition using reward-modulated spike-timing-dependent plasticity in deep convolutional networks," *Pattern Recognition*, vol. 94, pp. 87–95, 2019.

[20] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[21] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics*, vol. 1, no. 1, p. 52, 2018.

[22] P. Dudek, S. Szczepanski, and J. V. Hatfield, "A high-resolution cmos time-to-digital converter utilizing a vernier delay line," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, pp. 240–247, 2000.

[23] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, p. 075201, 2012.

[24] P.-Y. Chen and S. Yu, "Compact modeling of rram devices and its applications in 1t1r and 1s1r array design," *IEEE Transactions on Electron Devices*, vol. 62, no. 12, pp. 4022–4028, 2015.

[25] Z. Jiang, S. Yu, Y. Wu, J. H. Engel, X. Guan, and H.-S. P. Wong, "Verilog-a compact model for oxide-based resistive random access memory (rram)," in *2014 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. IEEE, 2014, pp. 41–44.

[26] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor–cmos system for efficient multiply–accumulate operations," *Nature Electronics*, vol. 2, no. 7, pp. 290–299, 2019.

Link-Level Abstraction of IEEE 802.11ay based on Quasi-Deterministic Channel Model from Measurements

Neeraj Varshney, Jiayi Zhang, Jian Wang, Anuraag Bodi, and Nada Golmie
Wireless Networks Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA
Emails: {neeraj.varshney, jiayi.zhang, jian.wang, anuraag.bodi, nada.golmie}@nist.gov

***Abstract*—In this paper, we analyze the performance of link-level abstraction for orthogonal frequency-division multiplexing (OFDM) and single-carrier (SC) modes in IEEE 802.11ay wireless systems over the 60 GHz millimeter-wave band. In particular, we evaluate the effectiveness of the three existing effective signal-to-noise ratio (SNR) metric (ESM) schemes (i.e., exponential ESM (EESM), mean mutual information per coded bit (MMIB) and post-processing ESM (PPESM)). Furthermore, to deal with the issue that EESM calibration is dominated by channel realizations with poor error performance, we introduce a classification based EESM (CEESM) scheme with a new metric named coefficient of variation, which is used to measure the severity of frequency-selective fading. Finally, we present several important insights developed through extensive experimentation. Based on our validation results, the MMIB and PPESM can be employed with minimum computational complexity for OFDM and SC modes, respectively. In contrast, EESM and CEESM can be considered for both modes with better accuracy, but at a cost of high implementation complexity.**

***Index Terms*—Link-level abstraction, IEEE 802.11ay, Quasi-Deterministic channel model, Millimeter-wave**

I. Introduction

IEEE 802.11ay is an enhanced wireless local access network (WLAN) standard that is capable of achieving high throughput and high power efficiency based on its predecessor, IEEE 802.11ad standard [1]. In IEEE 802.11ay, the wireless transceiver operates in the 60 GHz millimeter-wave (mmWave) band with the aid of directional antenna beams. During the development of the 802.11ay standards, several algorithms and solutions have been proposed to improve system performance. Thus, how to evaluate the system performance with high fidelity is critical. Particularly, the physical (PHY) layer is an essential component in wireless systems and its performance directly affects the overall system performance.

For an instantaneous channel realization, it is important to predict the link-level performance, such as packet error rate (PER) or packet success rate (PSR), for different PHY-layer configurations (e.g., modulation and coding schemes (MCS), and antenna configurations) so that whether a packet was transmitted successfully or not can be determined. As it is not feasible to run the time-consuming symbol-by-symbol link-level simulation (LLS) simultaneously with the system-level simulation (SLS), a PHY-layer abstraction model should be used to predict the decoding results based on the multi-path properties of fading channels. This motivates us to focus on assessing the effectiveness of methods that can be used to abstract the PHY-layer performance of IEEE 802.11ay over the 60 GHz mmWave channel.

Existing research efforts have been conducted on link-to-system (L2S) level mapping [2]–[8]. Most of these works [2]–[5] have focused on the orthogonal frequency-division multiplexing (OFDM) systems, while a few other works [6], [7] have emphasized single-carrier (SC) systems for the Long-Term Evolution (LTE) uplink, which are all operated in the sub-6 GHz spectrum. Recall that IEEE 802.11ay WLAN operates in the 60 GHz mmWave band, which has unique characteristics in comparison to sub-6 GHz channels, including sparse signal paths raised by higher penetration loss, weaker diffraction, and higher directional signals introduced by the beamforming technology [9], among others. Thus, in this paper, we address the following issues: *(i) whether existing L2S mapping approaches are still suitable for mmWave OFDM communications, (ii) whether existing L2S mapping approaches can be applied to IEEE 802.11ay SC mode as they have been used in orthogonal frequency-division multiple access (OFDMA) and SC-FDMA systems, and (iii) unlike previous work using PER, we consider bit error rate (BER), which offers flexibility in media access control (MAC) layer to derive PER if packets are partially overlapped in time domain or frequency domain.* To address the aforementioned issues, in this paper we investigate and compare different L2S abstraction schemes for IEEE 802.11ay single-input single-output (SISO) communication for both SC and OFDM modes as an initial step of our project. Once we gain enough confidence in single-link abstraction schemes, we will extend our work for multiple-input and multiple-output (MIMO) systems.

In this paper, our primary contributions are as follows: (i) *We develop a LLS by incorporating the Quasi-Deterministic (Q-D) model extracted from mobile empirical measurements.* We conduct an IEEE 802.11ay LLS for both SC and OFDM modes based on the Q-D channel model, which emulates a lecture room (LR) environment. (ii) *Based on LLS results, we evaluate the effectiveness of the three existing effective signal-to-noise ratio (SNR) metric (ESM) schemes*, i.e., exponential ESM (EESM) [5], mean mutual information per coded bit (MMIB) [2] and post-processing ESM (PPESM) [7]. For EESM, the LLS results are used to calibrate the parameters, which are used to map subband signal-to-noise ratios (SNRs) to the effective SNR. For MMIB and PPESM, the LLS is used to validate the effectiveness of the SNR mapping, since no calibration is required. Further, to deal with the issue that EESM calibration is dominated by channel realizations with poor BER performance, we introduce a classification based EESM (CEESM) scheme with a new metric named coefficient

Varshney, Neeraj; Zhang, Jiayi; Wang, Jian; Bodi, Anuraag; Golmie, Nada T. "Link-Level Abstraction of IEEE 802.11ay based on Quasi-Deterministic Channel Model from Measurements." Presented at 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall). October 04, 2020 - October 07, 2020.

of variation (CV), which is used to measure the severity of the frequency-selective fading. With CEESM, the prediction accuracy of the BER performance for an instantaneous channel realization can be improved. Further note that, to the best of our knowledge, this is the first work which evaluates the ESM schemes on 802.11ay systems based on Q-D channel model. (iii) *We conduct extensive experiments in three environments to validate the effectiveness of the aforementioned schemes.*

The remainder of the paper is organized as follows. In Section II, we introduce the system model and preliminaries including the Q-D channel model, LLS and subband SNRs. In Section III, we present several L2S mapping schemes for OFDM and SC modes in detail. In Section IV, we present the simulation results. Finally, we conclude the paper in Section V.

II. System Model and Preliminaries

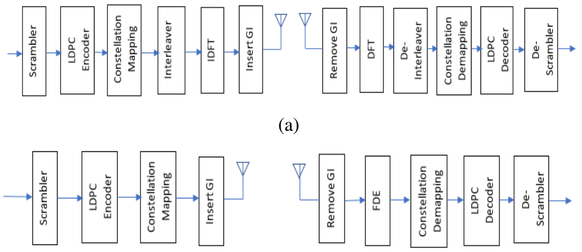
A. Q-D Channel Model

To model the channel, the two communication devices (i.e., transmitter (TX) and receiver (RX)) are randomly deployed in a LR of size 10 m × 19 m × 3 m. The phased-array-antenna with 5-by-5 antenna elements is used at both TX and RX. In our setup, we assume that the beam selection phase has been completed, and the optimum beam direction has been selected to achieve the maximum receive power. With a TX and RX pair, we use the Q-D channel realization software developed by the National Institute of Standards and Technology (NIST) to realize the channel [9]. The Q-D channel realization software has two major engines, a deterministic engine and a stochastic engine, to compute the rays between TX and RX. The deterministic engine generates deterministic rays, also referred to as specular rays using the ray-tracing method. On the other hand, the stochastic engine regenerates diffused rays, which are clustered around the specular rays using the Q-D parameters extracted from NIST measurement campaign [10]. In our realization, we only consider reflections up to 2nd order and each of the rays is characterized by the path gain, the delay τ, the angle-of-arrival (AOA) θ^R at the RX, and the angle-of-departure (AOD) θ^T from the TX. The generated Q-D rays are then analog beamformed by the TX and RX antenna arrays to generate the multipath between TX and RX. The detailed description about the Q-D channel model can be found in [9]. Based on this model, the impulse response of the beamformed channel is given as,

$$h(t) = \sum_{\tau} \sum_{\theta^T} \sum_{\theta^R} G^T(\theta^T) G^R(\theta^R) h(t, \tau, \theta^T, \theta^R), \quad (1)$$

where $G^T(\theta^T)$ and $G^R(\theta^R)$ represent the transmit and receive antenna beam patterns that weight each ray based on its direction. By combining all the rays falling into a sampling interval T_s, the channel is converted to a tap-delay profile, where the Kth tap is given as,

$$h(t, K) = \sum_{\tau=(K-1)T_s}^{KT_s} \sum_{\theta^T} \sum_{\theta^R} G^T(\theta^T) G^R(\theta^R) h(t, \tau, \theta^T, \theta^R). \quad (2)$$



(a)

(b)

Fig. 1: Transceiver diagrams of (a) OFDM mode and (b) SC mode

The channel impulse response $h(t, k) = \sum_K h(t, K)\delta(k - K)$ at the time t is subsequently used in LLS to obtain the fading channel performance. Note that t denotes the packet realization time in our simulation, and a new channel realization is generated for every packet to be transmitted.

B. Link-level Simulation (LLS)

The transceiver diagrams of the OFDM and SC modes are shown in Fig. 1(a) and Fig. 1(b), respectively. In this work, we assume that the perfect channel knowledge is available at the RX. Also, the PHY-layer impairments (carrier frequency offset, phase noise, etc.) are not included in the LLS. In the evaluation, the symbol-by-symbol simulation is performed. The packet size is considered as 4 096 bytes for both SC and OFDM modes. Note that in our simulation, the performance metric is the BER, which can be employed to compute the PER based on the transmit PSDU packet size. The end-to-end simulation is based on the 802.11ad implementation in the MATLAB WLAN Toolbox[1] with the extension to include 802.11ay specific MCSs. Moreover, we assume that the fading channel is semi-static and does not change within a packet transmission duration.

With the LLS, we obtain the BER performance as a function of SNR over additive white Gaussian noise (AWGN) channel and multi-path fading channels derived from the Q-D channel model for various MCSs. The AWGN performance can be stored in a network simulator such as ns3 so that BER can be directly retrieved based on mapped effective SNR. In the meantime, these decoding results obtained using the Q-D channel model for the LR environment, along with its instantaneous channel tap-delay profiles and the AWGN performance curve, will be used to calibrate the mapping parameters in EESM.

In addition, we use these fading channel decoding results to validate the effectiveness of SNR mapping by measuring the difference between the SNR of the AWGN channel and the mapped effective SNR given a channel realization for the same BER value. Note that the L2S mapping is to find the AWGN equivalent SNR (effective SNR) of the channel, and use the effective SNR to obtain the channel decoding performance (i.e., BER or PER) from the stored AWGN link performance

[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 NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

curves. Also, note that the calibration and validation are MCS and PHY mode dependent. The SC mode supports MCS 1 to MCS 21, and the modulation schemes are $\pi/2$-binary phase shift keying (BPSK), $\pi/2$-quadrature PSK (QPSK), $\pi/2$-16-quadrature amplitude modulation (QAM), and $\pi/2$-64-QAM for MCS 1 to MCS 6, MCS 7 to MCS 11, MCS 12 to MCS 16, and MCS 17 to MCS 21, respectively. Moreover, the OFDM mode supports BPSK, QPSK, 16-QAM, 64-QAM for MCS indices 1 to 5, 6 to 10, 11 to 15, and 16 to 20, respectively.

C. Subband SNRs

The subband SNRs are used to compute the effective SNR of a wide-band channel[2]. For this purpose, we first need to compute the center frequency of each subband and then derive the received subband SNRs in OFDM and SC Modes.

For a given wide-band channel of bandwidth B and its center frequency f_c, the center frequency of the nth subband can be obtained as, $f_{c,n} = f_c + n\Delta f$, where the subband spacing Δf considering the number of N_{ST} subbands is computed as, $\Delta f = \frac{B}{N_{\text{ST}}}$. Here $n = -\text{round}\left(\frac{N_{\text{ST}}}{2}\right) + 1, \cdots, \text{floor}\left(\frac{N_{\text{ST}}}{2}\right)$, where floor$(x)$ rounds the elements of x to the nearest integers towards $-\infty$ and round(x) rounds towards the nearest decimal or integer. In OFDM mode, each subband corresponds to a subcarrier with subcarrier spacing $\Delta f = 5.156\,25\,\text{MHz}$; while for SC mode, the entire bandwidth B is divided into $N_{\text{ST}} = 512$ subbands with $f_{c,n} = f_c + (n - \frac{1}{2})\Delta f, n = -(\frac{N_{\text{ST}}}{2} - 1), \cdots, \frac{N_{\text{ST}}}{2}$ to match the 512-point frequency-domain equalizer (FDE) adopted at the RX. The frequency-domain channel transfer function (CTF), which is defined as the Fourier transform of the impulse response $h(t)$ at subband center frequency $f_{c,n}$, can be obtained as, $H_{\text{BF}}(f_{c,n}) = \sum_{l=0}^{L-1} h(t,l)\exp(-j2\pi f_{c,n} l T_s)$, where T_s is the sampling period, l is the tap index, and L is the number of sampled multi-path. The received SNR for nth subband corresponding to the transmission over a SISO beamformed channel with power P can be derived by,

$$\gamma_n = \frac{P|H_{\text{BF}}(f_{c,n})|^2}{N_0}. \tag{3}$$

Here, the AWGN power N_0 is BkT for SC mode, where $k = 1.3807 \times 10^{-23}\,\text{J/K}$ is Boltzman constant, $T = 290$ is ambient temperature in degree Kelvin, and the bandwidth is $B = 2.16\,\text{GHz}$. For OFDM mode, $N_0 = NB_{\text{SB}}kT$, where the subcarrier bandwidth $B_{\text{SB}} = 5.156\,25\,\text{MHz}$ and $N = N_{\text{SD}} + N_{\text{SP}}$ denotes the total number of data and pilot subcarriers (i.e., $N = 336 + 16 = 352$) for a single $2.16\,\text{GHz}$ channel.

III. Link-to-System (L2S) Mapping Schemes

In order to evaluate the system wide performance for a wireless network with numerous transmission links and interference among multiple wireless nodes, the SLS requires the link-level performance in terms of the channel condition (e.g., signal-to-noise ratio and channel state information subject to multi-path fading) and system parameters, which provides the prediction of the instantaneous transmission qualities over specific individual links. L2S can provide such prediction without running through PHY-layer symbol-by-symbol simulation which can introduce very large simulation overhead.

Generally speaking, the L2S mapping is a class of methods to perform the PHY-layer abstraction and provide the prediction of error performance at a SNR in system-level. In the following sub-sections, we first introduce the effective SNR metric for the L2S mapping and subsequently present the EESM, MMIB, PPESM and CEESM schemes in detail.

A. Effective SNR Metric (ESM)

ESM provides the PHY-abstraction metric for L2S mapping which is an interface between the LLS and SLS [11]. Due to multi-path, the channel experiences frequency selective fading, meaning that channel frequency response is no longer flat and each frequency subband may have different gain. According to [12], for a given subband SNR vector of size N corresponding to a fading channel, the scalar effective SNR is defined as an equivalent SNR over AWGN channel, which would yield the same frame/ packet error probability. Note that the error performance curves over AWGN channel are MCS and PHY mode (i.e. SC or OFDM) dependent. For a given PHY mode and MCS, the mapping of the error performance between a specific fading channel and AWGN channel is represented by the AWGN equivalent SNR provided by EESM, MMIB, and PPESM. Based on the mapping function, the SLS can directly use the AWGN lookup tables to find the BER or PER corresponding to the data transmission through the multi-path fading channel.

Using a general L2S mapping technique with mapping function $\Phi(\cdot)$, the effective SNR γ_{eff} can be calculated as,

$$\gamma_{\text{eff}} = -\beta_1 \Phi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\Phi\left(-\frac{\gamma_n}{\beta_2}\right)\right), \tag{4}$$

where $\Phi^{-1}(\cdot)$ is the inverse mapping function of $\Phi(\cdot)$, and β_1 and β_2 are the scalar parameters which need to be optimized for different MCSs.

B. EESM

The EESM replaces the function $\Phi(\cdot)$ with the exponent function assuming that all the subcarriers are modulated using the same MCS. The mapping of EESM is given as,

$$\gamma_{\text{eff}} = -\beta \ln\left(\frac{1}{N}\sum_{n=1}^{N}\exp\left(-\frac{\gamma_n}{\beta}\right)\right), \tag{5}$$

where $\beta_1 = \beta_2 = \beta$ and $N = N_{SD}$ for OFDM mode. Similarly, we extend the EESM for SC mode with $N = N_{\text{ST}}$ subbands. In contrast to $N_{\text{SD}} = 336$ data subcarriers in OFDM mode, the above expression for SC mode considers all the $N_{\text{ST}} = 512$ subbands as the transmitted SC signal covers the entire bandwidth of $2.16\,\text{GHz}$. The tuning parameters β_{OFDM} and β_{SC} in Table I for 802.11ay OFDM and SC modes are calibrated based on the AWGN curves and the aggregated Q-D fading channel results with 600 random channel realizations and 161 SNR points for each realization. Note that this work

[2] Note that a single contiguous $2.16\,\text{GHz}$ channel with $f_c = 60.48\,\text{GHz}$ is considered in this study.

MCS Index	Optimal β_{OFDM}	MSE (dB)
1	1.60	-15.68
2	1.34	-14.43
3	1.29	-13.46
4	1.42	-12.67
5	1.25	-10.96
6	1.47	-11.07
7	1.75	-9.665
8	1.63	-8.477
9	1.82	-7.077
10	1.81	-6.038
11	3.89	-3.536
12	4.93	-1.884
13	5.37	-0.348
14	5.64	1.804
15	6.53	2.543
16	13.81	6.876
17	13.11	8.819
18	24.43	10.73
19	25.94	12.18
20	31.02	14.35

MCS Index	Optimal β_{SC}	MSE (dB)
1	0.21	-46.98
2	0.58	-42.44
3	0.87	-33.37
4	1.07	-33.18
5	1.31	-28.86
6	1.47	-26.38
7	1.15	-25.52
8	1.49	-23.27
9	1.82	-18.86
10	2.04	-16.98
11	2.24	-13.76
12	1.90	-9.030
13	2.92	-4.449
14	3.96	-0.241
15	4.71	1.781
16	6.68	4.352
17	4.49	1.290
18	7.37	6.094
19	11.51	10.97
20	14.33	13.65
21	20.50	16.27

TABLE I: Optimal β_{OFDM} and β_{SC} for OFDM and SC modes, respectively, which minimizes the mean-squared error (MSE) between AWGN SNR and predicted SNR vectors for same BER values.

combines all the channel realizations considered in CEESM while optimizing the tuning parameter in EESM.

C. MMIB based ESM

The MMIB based ESM was introduced for OFDM mode in IEEE 802.16m [2] and it was recently recommended in the evaluation methodology of IEEE 802.11ay standardization [13]. In MMIB, the function $\Phi(\cdot)$ maps each bit-channel to a mutual-information (MI) value, i.e., the capacity of the bit-channel [2]. The MMIB of the wide-band channel $\overline{I}_m$ is then computed by averaging the MIs of all bit-channels as, $\overline{I}_m = \frac{1}{N}\sum_{n=1}^{N} I_m(\gamma_n)$, where $N = N_{SD}$ for OFDM, $N = N_{ST}$ for SC, and m is the modulation order with value $m = \{1, 2, 4, 6\}$ representing BPSK, QPSK, 16-QAM, and 64-QAM, respectively. The term $I_m(\gamma_n)$ computes the MI per bit for nth subband with modulation order m as a function of nth subband SNR γ_n. Then, the reverse mapping function $I_m^{-1}(\overline{I}_m)$ is used to map the MI of the wide-band channel back to its effective SNR as described in [2, Section 3.2.1.5]. Since it is difficult to obtain the inverse function in a closed-form during implementation, especially for higher order modulations, we use a table-lookup method[3] to obtain the effective SNR from averaged MI.

D. Post-Processing ESM (PPESM) for SC-FDE

When communicating over a frequency-selective fading channel, both OFDM and SC modes employ FDE using various algorithms, i.e., matched-filtering (MF), zero-forcing (ZF), or the minimum MSE (MMSE), for post-processing at the RX. Further note that the channel encoding/decoding and symbol modulation/demodulation in OFDM mode are processed in frequency-domain directly; while these processes in SC mode are performed in time-domain. Thus, the OFDM modulated symbol can be recovered directly after one-tap equalization. However, in SC mode, the modulated symbols within a time-domain transmit block interfere with each other, which results in a residual interference among modulated symbols within a SC block after FDE operation, namely the inter-symbol-interference (ISI)[4]. Note that the residual ISI limits the system performance unless interference cancellation or decision-feedback equalization is introduced [14].

Considering the perfect CSI at the RX, the post-processing SNR of SC-FDE can be expressed as [7], $\gamma_{\text{SC}} = \frac{\mathcal{S}}{\mathcal{I}+\mathcal{N}}$, where the post-processing received signal power $\mathcal{S} = P\left|\frac{1}{N_{\text{ST}}}\sum_{n=1}^{N_{\text{ST}}} W_n^* H_n\right|^2$, the post-processing noise power $\mathcal{N} = \frac{N_0}{N_{\text{ST}}}\sum_{n=1}^{N_{ST}} |W_n|^2$, and the post-processing residual ISI power $\mathcal{I} = P\Big(\frac{1}{N_{\text{ST}}}\sum_{n=1}^{N_{\text{ST}}} |W_n^* H_n|^2 - \left|\frac{1}{N_{\text{ST}}}\sum_{n=1}^{N_{\text{ST}}} W_n^* H_n\right|^2\Big)$. Note that H_n and W_n represent the frequency-domain channel gain and corresponding FDE weight at the nth subband, respectively. Since the SC mode in our study employs a classic linear MMSE-FDE, its post-processing SNR can be derived as,

$$\gamma_{\text{SC}}=\left[\left(\frac{1}{N}\sum_{n=1}^{N}\frac{\gamma_n}{\gamma_n+1}\right)^{-1}-1\right]^{-1}=\left(\frac{1}{N_{\text{ST}}}\sum_{n=1}^{N_{\text{ST}}}\frac{1}{\gamma_n+1}\right)^{-1}-1, \quad (6)$$

where γ_n is given by Eq. (3). Consequently, the Eq. (6) can be expressed as the effective SNR γ_{eff} in Eq. (4), having a function of $\Phi(\gamma_n) = (\gamma_n + 1)^{-1}$. We refer to this scheme as PPESM for SC-FDE.

E. Channel Classification Based EESM

1) Classification of Channel Realizations: For 60 GHz mmWave-channel model, the performance of the system varies significantly between different TX and RX locations. Apart from this, different realizations for fixed locations of TX and RX also affect the performance drastically. If we use all the channel realizations while optimizing the parameters for mapping schemes (e.g. EESM), the system performance is always dominated by the worse channel realizations. Thus, the parameters optimized for various mapping schemes cannot be considered for realizations that experience frequency-flat fading.

To address this issue, we propose the channel classification based mapping scheme, in which channel realizations are grouped in different segments before obtaining the optimal parameters for mapping schemes. It is important to note that the random channel realizations can be easily divided into different segments using the metric such as 'CV'. The CV value using the N subcarrier channel gains, i.e., $|H_{\text{BF}}(f_{c,n})|, 1 \leq n \leq N$, for a given channel realization is computed as[5],

$$CV = \frac{std(|H_{\text{BF}}(f_{c,n})|, n = 1, 2, \cdots, N)}{mean(|H_{\text{BF}}(f_{c,n})|, n = 1, 2, \cdots, N)}, \quad (7)$$

[3] To apply the inverse function, we read the lookup table, and locate the closest SNR value for an input value of MI.

[4] The MF and MMSE based FDEs introduce the residual ISI. The ZF FDE completely eliminate the residual ISI, but the noise may be amplified when a deep frequency-domain fade is encountered.

[5] It is important to note that we consider subcarrier channel gains in CV to avoid the dependency on the transmit power while characterizing the channel realizations into different segments.

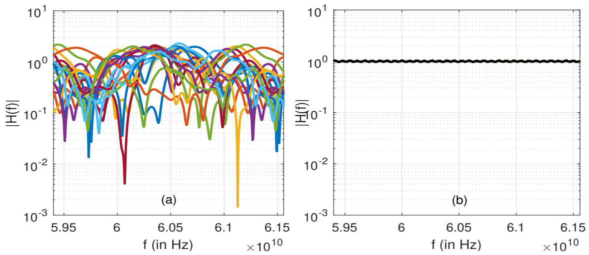


Fig. 2: CTF of randomly generated channel realizations for (a) segment 1 when CV > -2 dB, and (b) segment 6 when CV < -6 dB

where $std(\cdot)$ and $mean(\cdot)$ denote the standard deviation and mean, respectively. For a channel realization with high frequency-selective fading, the CV value would be very high in comparison to the frequency-flat fading channel. This can be clearly seen in Fig. 2, in which the channel realizations are divided into six segments using five thresholds with 1 dB spacing, i.e., $(-2, -3, -4, -5, -6)$ dB. From Fig. 2(a), we can observe that the segment 1 consists of all the channel realizations with $CV > -1$ dB and suffer from very high frequency-selective fading. In contrast to this, each realization in segment 6 with $CV < -6$ dB experiences almost frequency-flat fading, which can be clearly seen in Fig. 2(b).

2) Applying Classification to EESM: Using K thresholds, we first divide the channel realizations into $K+1$ segments and then calibrate the EESM parameter (β) using the channel realizations belonging to a particular segment only. By doing this, the prediction error can be significantly reduced, particularly when the channel experiences frequency-flat fading.

In order to map subband SNRs to the effective SNR of the channel, we first categorize the channel to a segment[6] based on the CV value as shown in Fig. 3. We then use the segment and MCS specific mapping function (i.e. specific β value) to perform the effective SNR mapping in EESM. The EESM parameters[7] corresponding to six segments for OFDM and SC modes are presented in Table II, respectively. It can be observed in Tables I and II that the optimal β values for EESM are approximately identical to the ones obtained for segment 1 in CEESM. This is owing to the fact that the system performance is always dominated by the worse channel realizations in EESM. Further, it is worth noting that the optimal β values for segment 6 are different than segment 1.

IV. Performance Evaluation

This section presents simulation results to validate the performance of various mapping methods over 60 GHz mmWave TGay channel model [15] developed by the NIST and MATLAB. We configure the MATLAB TGay channel model to have an access point (AP) with a 2-by-2 TX antenna array at a height of 6 m in an open area hotspot (OAH) environment, and an AP TX at 6 m in height with a 4-by-4 antenna array for a large hotel lobby (LHL) environment. In both OAH and LHL environments, the station (STA) has a RX height of 1.5 m. Since the antenna array sizes are not specified in [15], we configure the OAH and LHL environments using different antenna array sizes in order to simulate diverse beam patterns.

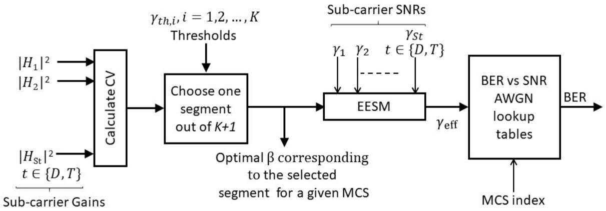

$\gamma_{th,i}, i = 1,2,\ldots,K$
Thresholds
Sub-carrier SNRs
γ_{St}
γ_1 γ_2 $t \in \{D,T\}$
$|H_1|^2$
$|H_2|^2$
$|H_{St}|^2$
$t \in \{D,T\}$
Sub-carrier Gains
Calculate CV
Choose one segment out of K+1
EESM
γ_{eff}
BER vs SNR AWGN lookup tables
BER
MCS index
Optimal β corresponding to the selected segment for a given MCS

Fig. 3: Schematic Block Diagram for CEESM

For simulation purposes, various system parameters are considered as follows. For both SC and OFDM modes, long guard interval length, i.e., 192 for OFDM and 128 for SC, is considered to reduce the interference between concatenated SC or OFDM transmission blocks[8], namely inter-block interference (IBI), especially for the scenario where the channel experiences large path delay. The PSDU packet length is set as 4 096 bytes, i.e, the number of bits transmitted within a packet (L_{bits}) is 32 768. It is worth noting that to implement the L2S mapping in SLS, the PER (P_{pe}) can be easily obtained from BER (P_{be}) as $P_{pe} = 1-(1-P_{be})^{L_{\text{bits}}}$ for any arbitrary packet size. Thus, in contrast to using PER-based lookup tables, the BER-based lookup tables can be used for any arbitrary packet size to decide its PSR. All the mapping schemes are validated over a total of 1 000 random channel realizations. For each packet transmission, the transmit signal power is randomly selected so that the receive signal power is among a range with a step size[9], plus a MCS-dependent offset[6].

From Fig. 4, it can be clearly seen that the predicted BER values obtained using the optimal β_k (c.f. Table II) in CEESM effective SNR match well with the actual BER for low as well as high MCSs in both OFDM and SC modes. It can also be observed in Fig. 4 that the optimal mapping parameters obtained through the NIST LR environment are also applicable to various MATLAB TGay-channel environments, including LHL and OAH. Using EESM and CEESM, the MSE between the AWGN SNR and predicted SNR vectors for the same BER values under various environments are given in Table III. It can be clearly seen that the CEESM performs well in comparison to the EESM.

Comparing to EESM and CEESM, MMIB and PPESM

[6] This work divide all the channel realizations into six segments using five thresholds and compute six EESM parameters for each MCS.

[7] To obtain these parameters, we considered 100 random channel realizations in each segment and simulated 161 SNR points for each random channel realization, having receive power ranging from -120 dBm to -80 dBm with a spacing of 0.25 dBm and a MCS dependent SNR offset. In SC mode, we consider SNR offsets of $(0, 5, 10, 15)$ dBm for MCS indices 1 to 6, 7 to 11, 12 to 16, and 17 to 21, respectively. In OFDM mode, we consider SNR offsets of $(0, 5, 10, 15)$ dBm for MCS indices 1 to 5, 6 to 10, 11 to 15, and 16 to 20, respectively.

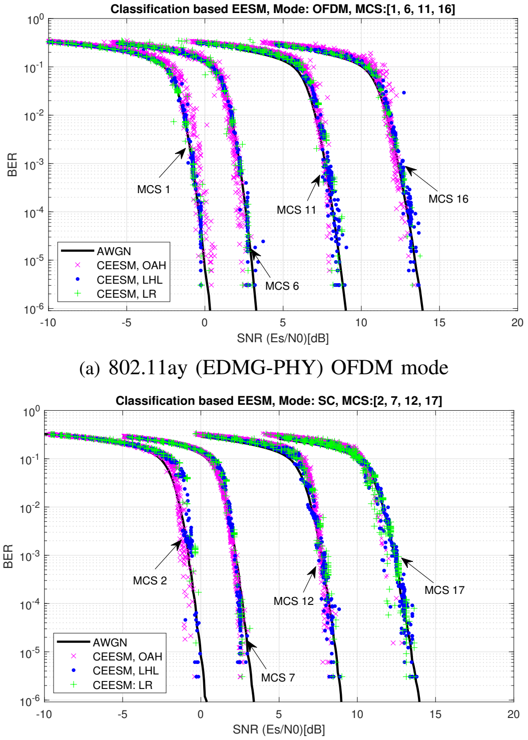
[8] We describe an OFDM symbol as an OFDM block in order to apply the same term to both OFDM and SC modes.

[9] The receive signal power range is given by: (i) EESM/CEESM: -120 dBm to -95 dBm with a step size of 0.5 dBm, and (ii) MMIB/PPESM: -120 dBm to -90 dBm with a step size of 0.05 dBm.

Varshney, Neeraj; Zhang, Jiayi; Wang, Jian; Bodi, Anuraag; Golmie, Nada T. "Link-Level Abstraction of IEEE 802.11ay based on Quasi-Deterministic Channel Model from Measurements." Presented at 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall). October 04, 2020 - October 07, 2020.

Mode	MCS Index	β_1	MSE	β_2	MSE	β_3	MSE	β_4	MSE	β_5	MSE	β_6	MSE
OFDM	1	1.63	-15.68	1.38	-16.77	1.68	-17.69	1.54	-17.69	2.09	-17.69	3.97	-18.86
	6	1.40	-10.86	1.35	-12.84	1.44	-13.01	1.81	-13.27	1.62	-14.09	4.98	-16.77
	11	4.78	-2.716	4.25	-3.269	3.92	-4.948	4.37	-5.436	3.26	-6.882	1.29	-10.75
	16	13.3	7.740	14.3	6.194	16.5	5.109	14.4	3.386	13.5	3.244	9.63	0.633
SC	2	0.83	-34.94	0.73	-35.85	0.72	-32.29	0.58	-30.22	0.51	-28.86	0.12	-34.55
	7	1.23	-23.09	1.24	-24.81	1.25	-24.20	1.24	-24.43	1.16	-24.09	0.49	-27.69
	12	2.51	-3.565	2.56	-5.482	2.21	-7.772	2.34	-8.601	2.31	-10.60	0.87	-19.58
	17	4.50	4.242	4.55	2.624	4.65	1.903	5.55	0.277	5.62	-3.251	5.37	-8.794

TABLE II: Optimal $\beta_{\mathrm{OFDM},k}, \beta_{\mathrm{SC},k} = \beta_k, 1 \leq k \leq 6$ corresponding to six segments for 802.11ay OFDM and SC modes



(a) 802.11ay (EDMG-PHY) OFDM mode

(b) 802.11ay (EDMG-PHY) SC mode

Fig. 4: Classification based EESM validation considering various MCSs (BPSK, QPSK, 16-QAM, 64-QAM all in $1/2$ code rate) with various 60 GHz mmWave multi-path fading channel models

<table>
<tr><th rowspan="2">MCS
Index</th><th colspan="3">CEESM OFDM</th><th colspan="3">EESM OFDM</th></tr>
<tr><th>LR</th><th>OAH</th><th>LH</th><th>LR</th><th>OAH</th><th>LHLL</th></tr>
<tr><td>1</td><td>-31.35</td><td>-24.82</td><td>-32.12</td><td>-29.16</td><td>-23.05</td><td>-30.31</td></tr>
<tr><td>6</td><td>-24.01</td><td>-22.18</td><td>-25.95</td><td>-22.79</td><td>-21.99</td><td>-24.90</td></tr>
<tr><td>11</td><td>-6.97</td><td>-6.09</td><td>-9.68</td><td>-5.65</td><td>-4.89</td><td>-7.10</td></tr>
<tr><td>16</td><td>1.32</td><td>9.85</td><td>-2.79</td><td>2.39</td><td>11.30</td><td>-1.50</td></tr>
<tr><th rowspan="2">MCS
Index</th><th colspan="3">CEESM SC</th><th colspan="3">EESM SC</th></tr>
<tr><th>LR</th><th>OAH</th><th>LHL</th><th>LR</th><th>OAH</th><th>LHL</th></tr>
<tr><td>2</td><td>-21.10</td><td>-23.20</td><td>-20.19</td><td>-17.64</td><td>-19.93</td><td>-17.91</td></tr>
<tr><td>7</td><td>-19.72</td><td>-24.51</td><td>-20.36</td><td>-18.47</td><td>-22.54</td><td>-19.43</td></tr>
<tr><td>12</td><td>-7.34</td><td>-4.81</td><td>-10.25</td><td>-5.72</td><td>-2.64</td><td>-7.89</td></tr>
<tr><td>17</td><td>1.13</td><td>5.15</td><td>0.48</td><td>2.80</td><td>6.56</td><td>0.72</td></tr>
</table>

TABLE III: MSE values (in dB) of CEESM and EESM schemes for OFDM/ SC modes under various TGay environments

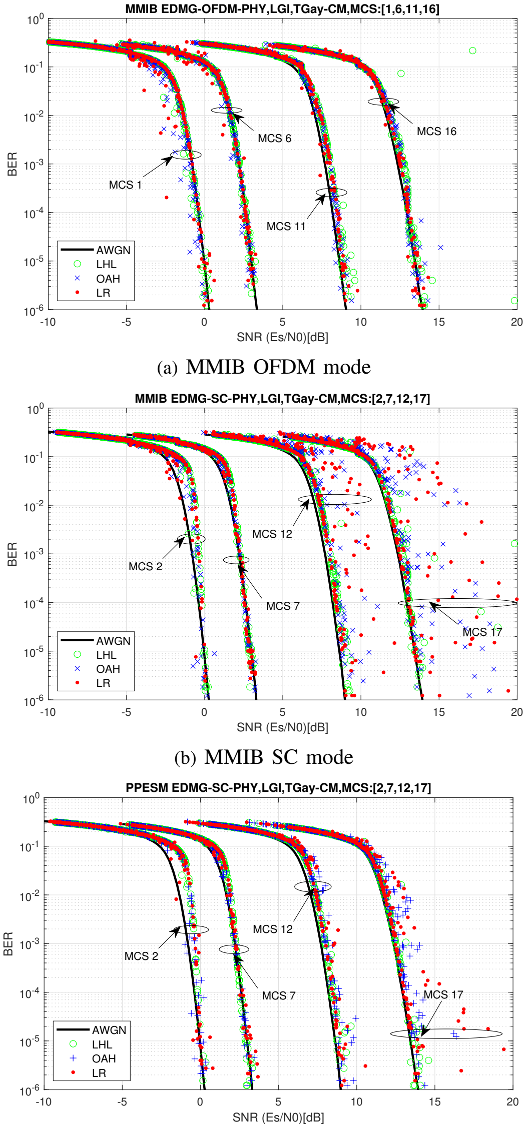
are generic and do not require any calibration. Figs. 5(a) and 5(b) show the BER/SNR validation results of MMIB scheme for OFDM and SC modes in various mmWave channel environments, i. e. NIST TGay LR, Matlab TGay LHL, and OAH. In Fig. 5(a), the predicted BER samples of OFDM mode using MMIB over channel realizations of all three environments closely match to the BER obtained through LLS over AWGN channel. As we observe that a few outliers exist for higher MCSs, the OFDM packet transmission over Matlab TGay LHL channel occasionally result in a high error rate despite of high SNR, as shown in Fig. 5(a). These few outliers in turn results in the high MSE values in Table IV[10]. The reason is that LHL is a reflection-rich environment. When a 2-by-2 antenna array is employed with wider half-power beam width (HPBW) in LHL environment, compared to a 4-by-4 antenna array in LR environment, some paths with strong power could have large delays, which makes the guard interval not sufficient to handle IBI for both OFDM and SC modes. In contrast to OFDM mode, the BER predicted using MMIB for SC mode does not match well with the LLS results for all three TGay channels when high order MCSs are employed, as shown in Fig. 5(b). The reason is that apart from the above-mentioned LHL environment issue, the SC mode is subject to the post-processing residual ISI at the RX with MMSE-FDE even when the path delays are within the guard interval length, which is not taken into account while computing the effective SNR in MMIB. Therefore, the residual ISI increases (or decreases) when the signal power increases (or decreases).

Fig. 5(c) demonstrates the BER/SNR validation comparison of PPESM for SC mode transmission in LR, OAH and LHL environments. It can be observed that the PPESM BER samples are closer to the AWGN curves when comparing with the MMIB results in Fig. 5(b), especially for LHL environment. This is owing to the fact that the PPESM is aware of the residual ISI while computing the effective SNR through post-processing SNR (c.f. (6)). However, some gap still exist between the predicted BER using PPESM and the BER from LLS for high order MCS, since the high order modulation schemes are more sensitive to the residual ISI, which in turn generates the distortion. Further note that this distortion is hard to remove unless employing a more advanced equalizer with a cost of higher complexity.

V. Final Remarks

In this paper, we have addressed the L2S level mapping for OFDM and SC modes in IEEE 802.11ay WLAN systems

[10] The MSE values 18.89 dB and 26.53 dB for MMIB OFDM MCS 11 and MCS 16 are obtained when the IBI is considered with the presence of corresponding channel realizations. Otherwise, the MSE values -6.19 dB and 1.02 dB are calculated with the absence of these realizations causing IBI.



(a) MMIB OFDM mode

(b) MMIB SC mode

(c) PPESM SC mode

Fig. 5: MMIB and PPESM validation considering various MCSs (BPSK, QPSK, 16-QAM, 64-QAM all in $1/2$ code rate)

MCS	MMIB OFDM		
Index	LR	OAH	LHL
1	-24.34	-24.49	-30.34
6	-24.15	-25.27	-28.38
11	-6.60	-8.54	18.89 (-6.19)
16	-1.72	3.21	26.53 (1.02)

MCS	MMIB SC			PPESM SC		
Index	LR	OAH	LHL	LR	OAH	LHL
2	-18.79	-19.52	-18.76	-18.63	-19.94	-18.04
7	-20.20	-21.46	-19.78	-19.81	-21.32	-20.69
12	9.41	3.11	18.88	-1.02	-2.72	-5.82
17	24.48	21.41	16.25	21.25	10.00	1.81

TABLE IV: MSE values (in dB)[9] of MMIB for OFDM/ SC modes and PPESM for SC Mode in various TGay environments

over the 60 GHz mmWave band using the Q-D channel model generated based on real measurements. We have studied the performance of the three existing ESM schemes (i.e., EESM, MMIB, PPESM) which were originally proposed for OFDMA and SC-FDMA systems and on sub-6 GHz bands. We have also designed the CEESM scheme by incorporating a metric to measure the severity of frequency-selective fading. To validate the feasibility of these schemes, we have conducted extensive simulations. Our experimental results confirm that EESM and CEESM achieve the highest accuracy but with the initial calibration cost. We compare the performance for three environments (LR, OAH and LHL), and confirm that the mapping parameters calibrated using one environment can be applied to other environments, so that the mapping metric is more MCS and mode dependent rather than environment dependent. We also confirm that MMIB and PPESM, having very low implementation cost, provide reasonable performance for OFDM mode and SC mode, respectively. With the initial effort completed for the single link systems, we are currently working to extend this research to evaluate MIMO systems.

References

[1] *Wireless LAN MAC and PHY Specifications–Amendment: Enhanced Throughput for Operation in License-Exempt Bands Above 45 GHz*, IEEE Std. IEEE P802.11ay/D4.0, June 2019.

[2] *IEEE 802.16m Evaluation Methodology*, IEEE Std. IEEE 802.16m-08/004r5, 2009.

[3] T. L. Jensen, S. Kant, J. Wehinger, and B. H. Fleury, "Fast link adaptation for MIMO OFDM," *IEEE Trans. Veh. Technol.*, vol. 59, no. 8, pp. 3766–3778, Oct 2010.

[4] F. L. Aguilar, G. R. Cidre, and J. R. París, "Effective SNR mapping algorithms for link prediction model in 802.16e," in *Proc. ICUMTW 2009*, Oct 2009, pp. 1–6.

[5] H. Liu, L. Cai, H. Yang, and D. Li, "EESM based link error prediction for adaptive MIMO-OFDM system," in *Proc. IEEE VTC2007-Spring*, April 2007, pp. 559–563.

[6] F. Jiang, Y. Xue, B. Jiang, and S. Jin, "Average effective SNR mapping in LTE-A uplink," in *Proc. IEEE ICCC 2012*, Aug 2012, pp. 403–407.

[7] Samsung, "3GPP TSG-RAN WG1 meeting #43: Simulation methodology for EUTRA uplink: SC-FDMA and OFDMA," 3GPP, Tech. Rep. Work Item R1-051352, Nov. 2005.

[8] Motorola, "3GPP TSG-RAN WG1 meeting #43: Simulation methodology for EUTRA UL: IFDMA and DFT-Spread-OFDMA," 3GPP, Tech. Rep. R1-051335, Nov. 2005.

[9] A. Bodi, J. Zhang, J. Wang, and C. Gentile, "Physical-layer analysis of IEEE 802.11ay based on a fading channel model from mobile measurements," in *Proc. IEEE ICC 2019*, May 2019, pp. 1–7.

[10] C. Lai and et al., "Methodology for multipath-component tracking in millimeter-wave channel modeling," *IEEE Trans. Antennas Propag.*, vol. 67, no. 3, pp. 1826–1836, March 2019.

[11] K. Brueninghaus and et al., "Link performance models for system level simulations of broadband radio access systems," in *Proc. IEEE PIMRC 2005*, vol. 4, Sep. 2005, pp. 2306–2311 Vol. 4.

[12] S. Nanda and K. M. Rege, "Frame error rates for convolutional codes on fading channels and the concept of effective Eb/N0," *IEEE Trans. Veh. Technol.*, vol. 47, no. 4, pp. 1245–1250, Nov 1998.

[13] *IEEE P802.11 Wireless LANs TGay Evaluation Methodology*, IEEE Std. IEEE 802.11-09/0296r16, 2016.

[14] J. Zhang, L. Yang, and L. Hanzo, "Frequency-domain turbo equalisation in coded SC-FDMA systems: EXIT chart analysis and performance," in *Proc. IEEE VTC-Fall 2012*, Sep. 2012, pp. 1–5.

[15] *Channel Models for IEEE 802.11ay*, IEEE Std. IEEE 802.11-15/1150r9, 2016.

Varshney, Neeraj; Zhang, Jiayi; Wang, Jian; Bodi, Anuraag; Golmie, Nada T. "Link-Level Abstraction of IEEE 802.11ay based on Quasi-Deterministic Channel Model from Measurements." Presented at 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall). October 04, 2020 - October 07, 2020.

NIST SP 1283
August 2022

Effects of shield gas flow on meltpool variability and signature in scanned laser melting

David C. Deisenroth[1], Jorge Neira, Jordan Weaver, and Ho Yeung
National Institute of Standards and Technology[2]
Gaithersburg, MD

Author copy Accepted to Proceedings of the ASME 2020 15th International Manufacturing Science and Engineering Conference 2020, June 22-26, 2020, Cincinnati, OH, USA

Abstract

In laser powder bed fusion metal additive manufacturing, insufficient shield gas flow allows accumulation of condensate and ejecta above the build plane and in the beam path. These process byproducts are associated with beam obstruction, attenuation, and thermal lensing, which then lead to lack of fusion and other defects. Furthermore, lack of gas flow can allow excessive amounts of ejecta to redeposit onto the build surface or powder bed, causing further part defects. The current investigation was a preliminary study on how gas flow velocity and direction affect laser delivery to a bare substrate of Nickel Alloy 625 (IN625) in the National Institute of Standards and Technology (NIST) Additive Manufacturing Metrology Testbed (AMMT). Melt tracks were formed under several gas flow speeds, gas flow directions, and energy densities. The tracks were then cross-sectioned and measured. The melt track aspect ratio and aspect ratio coefficient of variation (CV) were reported as a function of gas flow speed and direction. It was found that a mean gas flow velocity of 6.7 m/s from a nozzle 6.35 mm in diameter was sufficient to reduce meltpool aspect ratio CV to less than 15 %. Real-time inline hotspot area and its CV were evaluated as a process monitoring signature for identifying poor laser delivery due to inadequate gas flow. It was found that inline hotspot size could be used to distinguish between conduction mode and transition mode processes, but became diminishingly sensitive as applied laser energy density increased toward keyhole mode. Increased hotspot size CV (associated with inadequate gas flow) was associated with an increased meltpool aspect ratio CV. Finally, it was found that use of the inline hotspot CV showed a bias toward higher CV values when the laser was scanned nominally toward the gas flow, which indicates that this bias must be considered in order to use hotspot area CV as a process monitoring signature. This study concludes that gas flow speed and direction have important ramifications for both laser delivery and process monitoring.

Keywords: metal additive manufacturing, laser powder bed fusion, shield gas flow, plume, condensate, process monitoring

Nomenclature

AR	melt-track aspect ratio
CV	coefficient of variation
HS	hotspot
d	melt-track depth, m
w	melt-track width, m
μ	mean
σ	population standard deviation

1. Introduction

The importance of optimizing shield gas flow is evident from the continuously improving gas flow provisions in industrial laser powder bed fusion machines (LPBF). Inadequate shield gas flow is associated with porosity, lack of fusion defects, decreased strength, increased variability in mechanical properties, and high dimensional inaccuracy in components constructed with LPBF [1–5]. The following sections of this introduction reviews the literature on how laser melting byproducts can affect beam delivery and cause adverse

[1] Contact author: david.deisenroth@nist.gov

[2] 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.

effects on laser melting. Process monitoring approaches that may be used to detect adverse effects on melting will then be discussed. Finally, the objective of the current investigation will be described.

1.1 Laser Melting Byproducts

The properties of shield gas play an important role in both laser welding and in LPBF. Types of gas that are used in both laser melting processes include N, Ne, and Ar, and each gas can affect build quality with a given build material [3]. The gas solubility/reactivity with molten metal, and the wettability of the metal in a given gas environment affects keyhole stability and the resulting porosity in laser welding [6]. Finally, gas properties including density, thermal conductivity, and reactivity with metal vapor affect plume generation and persistence [7].

As stated previously, a gas environment that limits reactivity with the metal is insufficient to ensure consistent melting. The need for directional gas flow across the laser-metal interaction stems from the fact that power levels, velocities, and laser focus spot size commonly used for LPBF result in a significant amount of vapor generation at the laser-metal interaction area [8]. This vapor jet, through several mechanisms, produces byproducts that interfere with laser beam delivery. This interference in turn results in insufficient and inconsistent melting of the buildup material, and therefore defects.

The hot vapor jet does not appear to be, in itself, a direct source of beam interference. As the beam passes through the hot vapor, the vapor may become weakly ionized, but inverse Bremsstrahlung absorption due to plasma appears to be small compared to other byproduct effects in metal melting with fiber lasers [9,10]. Beam obstruction due to the metal vapor becomes significant as the vapor condenses in a cloud above the meltpool. When its primary constituent is iron, the metal vapor forms ultrafine condensate particles of 80 nm to 100 nm in diameter with number density of about 10^{10} particles per cm^3 [10]. These clouds of suspended condensate are associated with beam scattering and may attenuate the laser power in excess of 10 % when welding steels with a fiber laser [10]. Furthermore, the hot vapor jet causes localized refractive index changes that may defocus and/or redirect the beam, causing poorly controlled melting [11].

The hot vapor jet also contributes to generation of ejecta that range from 10 μm to 100 μm in diameter via several mechanisms. The high pressure and velocity of the vapor jet and the high surface temperature gradients driving Marangoni flows in the meltpool leads to rapid convection and oscillations that may result in droplet ejection from the meltpool [8]. Beam power oscillations due to condensate obstruction may occur at frequencies up to several kilohertz, which has also been shown to increase liquid ejecta (spatter) in laser welding [10].

In addition to being a driver of meltpool flows, the vapor jet also entrains the ambient gas around it, which may further impel nearby powder particles [12]. The powder can be drawn from the powder-bed from several meltpool widths away from the laser incidence location [8]. The solid metal powder particles, typically of diameter of about 60 μm, may be driven by cold ambient gas convection alone, or they may become entrained directly in the hot vapor jet [8].

Regardless of how they are generated or driven, ejecta of 10 μm to 100 μm in diameter that intersect the beam path may obstruct beam delivery and reduce local melt efficacy. Moreover, this material redistribution from the laser scan path to other areas on the part and powder bed are associated with balling and lack of fusion defects [13]. Finally, material redistribution may interfere with layer recoating, leading to further part defects [14].

Directional gas flow across the laser-metal interaction improves build quality by transporting nanoscale condensate, which reduces beam scattering and attenuation. Directional gas flow also transports microscale ejecta away from the beam path, built part, and powder bed, reducing beam disturbances and defects associated with redistributed material. Finally, directional gas flow also improves the accuracy of beam delivery by reducing thermal lensing by transporting hot gas and vapor from the beam path.

The most practical implementation of directional gas flow is with a single inlet and outlet in the build chamber, as illustrated by the gas flow design of common industrial LPBF machines. This configuration means that the gas flows one direction across the build plane under all conditions. Studies have shown that in LPBF, laser scanning along the direction of the gas flow is associated with increased scanning of airborne ejecta and decreased part tensile strength [14,15]. Scanning along the direction of gas flow is also associated with increased beam interference and particle emissions from suspended condensate in laser welding [9]. Scanning nominally against or perpendicular to the gas flow direction may alleviate these issues [9,14,15]. Similarly, gas flow direction should be considered when developing the hatching directions in scan strategies in order to avoid scanning through suspended condensate or ejecta that were generated by preceding scan tracks—i.e. hatching direction should be nominally toward or perpendicular to gas flow.

1.2 Melt Tracks

In the current investigation, cross-sectioning and measurement of the melt tracks was used as a method to evaluate whether sufficient gas flow was applied to remove byproducts from the beam path. Cross-sectioning and selective etching of solidified melt tracks allow for (destructive) analysis of the meltpool. With this approach, the mode of melting imposed on the substrate, grain structure, and mechanical properties of the solidified material can be evaluated [16,17]. Furthermore, melt track cross-sections are a comparison approach for validating multiphysics modeling results with experimental measurements [18].

The energy density applied to the material has a strong effect on the depth and width of the meltpool, and therefore if significant attenuation or distortion of the beam occurred, that effect was evidenced by the dimensions of the melt tracks. There are varying definitions of energy density reported in terms of energy per length, energy per volume, etc., but for the purposes of the current discussion, a strict definition of energy density is not necessary. 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 [18]. Beam scattering, attenuation, and lensing may widen the beam, forming a shallower and wider meltpool [11,13].

The energy density applied to the material is also related to the rate of vapor generation from the process. At process conditions producing little to no vapor jet, "conduction mode" occurs. Conduction mode is associated with small meltpool width and depth and an aspect ratio (depth divided by width) of less than unity. Among other defects, too low of energy densities are associated with lack of fusion and balling defects in LPBF [19]. 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 (AR) 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 the deep cavity formed in the meltpool by the vapor recoil pressure [20–22]. Among other defects, excessive or unstable keyholing is associated with residual porosity and loss of volatile alloy components [6,15]. In the current investigation, meltpool cross-sections will be used to measure the meltpool mode and its variability due to laser delivery with differing gas flow velocities and directions.

1.3 Process Monitoring

Real-time monitoring of the build process to determine part quality is preferred as an alternative to destructive evaluation to assess part quality. Process monitoring approaches have been suggested for assessing homogeneity of the powder bed, surface quality of as built surfaces, and part dimensions during the build [5]. The focus of the current investigation is in assessing the state and consistency of the meltpool in situ as a function of gas flow speed and direction.

A variety of process monitoring approaches relating to meltpool state and consistency have been reported in the literature. Photodetectors mounted on the chamber walls can be used to indicate poor absorption events that are associated with lack of fusion defects [1,23]. High-speed thermal imaging of process spatter and of the hot vapor plume have been investigated as a meltpool monitoring approaches for laser powder bed fusion as well as laser welding [23–25]. The current investigation uses inline imaging of the laser-induced hotspot. The hotspot image processing approach determines the hotspot size and variability. The in-situ hotspot size and variability are then compared with destructively-evaluated meltpool size and variability.

1.4 Objective of the Current Investigation

An example of tracks melted in Nickel Alloy 625 (IN625) with adequate and inadequate gas flow are shown in Figure 1a and Figure 1b, respectively. Both sets of tracks were melted in an argon environment, but one set with adequate gas flow and one with no directional gas flow across the laser-metal interaction area. The tracks with adequate gas flow shown in Figure 1a are highly consistent in width and texture along the length of tracks and from track to track, including the size of the terminal meltpool on the right end of each. In contrast, the tracks with no gas flow shown in Figure 1b are highly inconsistent in width and texture, both along length of tracks and from track to track.

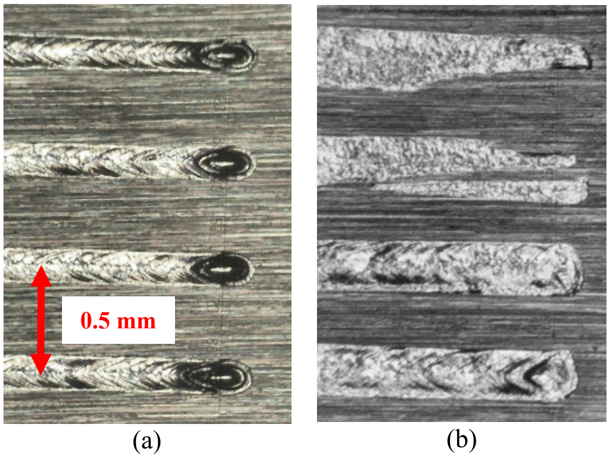


Figure 1: surface view of tracks laser melted with (a) adequate shield gas flow and (b) no shield gas flow. Both tracks were melted in an argon environment on inconel 625 with a 1070 nm fiber laser, a laser spot D4σ of 100 µm, a scan speed of 400 mm/s, and laser power 150 W. Images are approximately equivalent scale.

And so, it is evident both from the literature and the example shown in Figure 1 that laser melting in an inert gas atmosphere is insufficient, and that a directional gas flow across the laser-metal interaction area is essential to promote consistent melting and ensure final part quality in LPBF. The first goal of the current investigation is to better quantify the meltpool dimensions, AR, and AR variability generated when melting of IN625 in adequate and inadequate gas flow conditions. The second goal of this investigation was to relate melt track cross-sectional AR and AR coefficient of variation (CV) to inline process monitoring hotspot size and hotspot size variability in adequate and inadequate gas flow conditions. The process monitoring effort was undertaken to develop a tool that may identify adverse gas flow effects without destructive evaluation of built parts.

2. Materials and methods

The experiments reported here were performed in the National Institute of Standards and Technology (NIST) Additive Manufacturing Metrology Testbed (AMMT) [26]. 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, the working material was rolled and annealed bare plates of Nickel Alloy 625 (IN625).[3]

2.1 Hotspot Monitoring

The hotspot was imaged with an inline high-speed camera coaxially aligned with the heating laser. Emitted light from the meltpool is filtered by a bandpass filter at 850 nm (40 nm bandwidth at full-width half maximum) and diverted by a dichroic mirror to the camera sensor with nominal 1:1 magnification and 8 μm/pixel size. The images were taken with 45 μs exposure time, 120-pixel x 120-pixel window, and 8-bit grayscale. The grayscale level corresponding to the melting point is roughly determined by comparing the meltpool image with the physical melt-track. This inline meltpool images, therefore, consisted of an approximately circular hotspot with a radiance temperature ranging from approximately 1650 °C to 2050 °C. The camera is triggered by scan position, so the location where each image was taken is known. The equivalent frame rate is 20 kHz.

2.2 Oblique View of Meltpool and Plume

In order to visualize the relationship between the inline meltpool image and the process plume, a second high-speed camera was outfitted with a macroscopic view lens and aligned to obliquely view the meltpool that was scanned perpendicular to the view field. The oblique view camera was outfitted with a spectral filter to the range of approximately 830 nm to 870 nm, which is a similar range to the inline process monitoring camera. The inline and oblique view camera images were synchronized after data was collected.

2.3 Gas Flow Conditions

Gas flow was provided to the process plane at an angle illustrated in Figure 2. Using the coordinate system detailed in Figure 2, the gas flow tube axis was -31° from xz plane and -25° from xy plane. The unit vector formed by the nozzle orientation was $\hat{u} = -0.78\hat{x} - 0.47\hat{y} - 0.42\hat{z}$. The nozzle internal diameter was 6.35 mm. The center of tube outlet was approximately 9 mm in z direction from the surface and approximately 17 mm from the center of the origin shown in Figure 2. The gas flow speed was controlled with a rotameter with average nozzle outlet velocities of 0 m/s, 2.2 m/s, 6.7 m/s, and 22 m/s.

2.4 Scan Strategy

As shown in Figure 2, the IN625 substrate was scanned in six subsequent patches, sequentially as they are numbered. Patches 1 and 2 were scanned with 195 W at 800 mm/s, patch 3 and 4 were scanned with 150 W at 400 mm/s, and patch 5 and 6 were scanned with 195 W at 1200 mm/s. These process parameters are equivalent to those used in the 2018 Additive Manufacturing Benchmark Test Series (AM-Bench) and have very well characterized outcomes [27]. In patches 1, 3, and 5, the scan direction was from top to bottom and tracks were sequentially left to right. In patches 2, 4, and 6, the scan direction was from left to right and tracks were sequentially top to bottom. In other words, odd numbered patches were scanned nominally perpendicular to the gas flow with hatching direction nominally toward the flow and even numbered patches were scanned nominally toward the gas flow with hatching direction nominally perpendicular the flow. This approach facilitated two orthogonal gas flow directions relative to the scan direction. Hatch spacing of 0.5 mm was used to avoid heat accumulation and metallurgical interaction between the tracks. The lowest energy density was applied to

[3] 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.

patches 5 and 6, medium energy density was applied to patches 1 and 2, and highest energy density was applied to patches 3 and 4. The D4σ laser spot size used in all experiments was 100 µm.

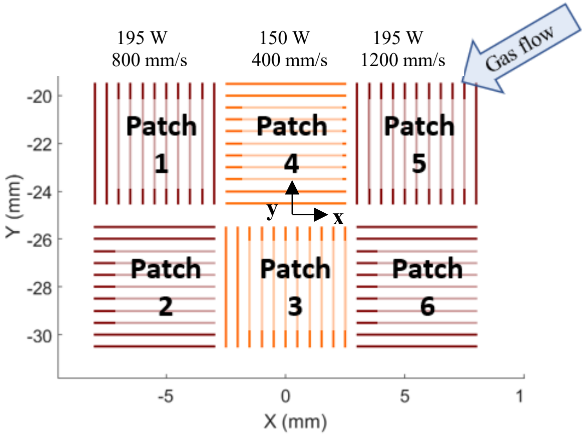


Figure 2: top view of scan tracks applied to the substrate. Coordinates are shown with the origin in the center and the z-axis positive direction out of the image plane.

2.4 Cross-Sectioning

The Ra roughness (defined by ISO 4287:1997, [28]) of the IN625 substrate in which tracks were melted was 19.8 µm with standard deviation between tracks of 3.4 µm. After laser scanning, the substrates were cross-sectioned orthogonal to the scan direction. The sections were 1 mm to 2.5 mm from either end of the melt tracks. Because of the spacing between patches, between 9 and 11 tracks from each patch were sectioned and measured. This sample of tracks from each patch was taken as a random sample of the track cross sections generated by each gas flow speed, scan speed, scan direction, and laser power combination.

Once cut, the substrate samples were hot mounted in mineral filled epoxy thermoset resin, then ground and polished to a mirror-like finish. Next, the mounted samples were etched with aqua regia to enhance contrast between the melt tracks and wrought material. The melt tracks were then optically imaged with 0.155 µm per pixel resolution. An example of a meltpool cross-section image from this work is shown in Figure 3. The meltpool width was measured from the substrate surface intersection with the meltpool edges and the depth was measured from the substrate surface to the deepest point of the meltpool with a pixel-based bounding box.

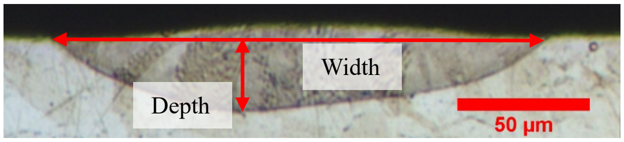


Figure 3: example image of melt track cross-section image showing locations from which width and depth were measured.

From the width (w) and depth (d) measurements, the aspect ratio (AR) of each meltpool was calculated with Equation (1).

$$\mathrm{AR} = \frac{\mathrm{d}}{\mathrm{w}} \tag{1}$$

Coefficients of variation were calculated for both the meltpool AR and the hotspot area as a measure of how much variability there was from the mean value at each combination of gas flow speed, scan speed, scan direction, and laser power. The coefficient of variation (CV) is simply the standard deviation of the population (σ) divided by the mean (µ) expressed as a percentage, as shown in Equation (2).

$$\mathrm{CV} = \frac{\sigma}{\mu} \times 100 \tag{2}$$

2.4 Uncertainty

The laser power of the AMMT was calibrated with a commercial power meter with reported accuracy of 2.5 % and 0.5 % repeatability and is, therefore, assigned 2.6 % uncertainty with 1σ confidence. The spot size has been measured and reported with two independent methods, and has been shown to be within 10 µm for all measurements taken at 100 µm spot size; Zhirnov et al. [29] compared measurements of the AMMT spot size with both low duty cycle attenuated laser power on a common camera, as well as a dynamic full-power beam sampler. Spot size is therefore assigned a 5 µm uncertainty with 1σ confidence.

The gas flow velocity was set with a manual rotameter that had an uncertainty of approximately 10 %. The orientation of the gas flow nozzle was estimated to be within 10° of the measured value. Gas flow speed and direction accuracy will be improved in future experiments.

Preliminary experiments have found that the uncertainty of change in position on the AMMT laser spot were within 0.25 % and is therefore assigned that value with 1σ confidence. Finally, it was estimated that the subjectivity uncertainty associated with the pixel-based bounding box used to measure meltpool width and depth results in a 67 % likelihood that the meltpool boundary lies within ±5 pixels of the assigned location. Therefore, with a pixel scale 0.155 µm per pixel, the measurement uncertainty is 0.8 µm with 1σ confidence. The uncertainty is summarized in Table 1.

Table 1: uncertainty of measured values with confidence intervals

Measurand	**Uncertainty**	**Confidence**
Laser power	2.6 %	1σ
Spot size	5 µm	1σ
Gas flow speed	10 %	Conservative estimate
Gas flow direction (x, y, and z)	5°	Conservative estimate
Track length	0.25 %	1σ
Meltpool width	0.8 µm	1σ
Meltpool depth	0.8 µm	1σ

3. Results and discussion

The experimental results are presented in this section. First, examples of the melt track cross-sections generated under adequate and inadequate gas flow are shown. The AR and AR CV of the tracks as a function of gas flow rate are then presented, then are used as a criteria to define adequate gas flow. Next, the obstructing effect of lack of gas flow on laser delivery is shown with synchronized inline hotspot images and an oblique view of the meltpool and hot vapor plume. After that, the inline hotspot area and its CV are presented as a function of gas flow speed. The relationships between hotspot area, hotspot area CV, meltpool AR, and meltpool AR CV, are finally presented.

3.1 Effects of Gas Flow Speed on Meltpool Cross-Section

This section will first qualitatively, then quantitatively, describe the effects of gas flow speed on meltpool AR and AR CV. Three laser energy densities were applied to the sample, and are shown in increasing order in Figure 4. The lowest energy density was applied in patches 5 and 6, medium energy density was applied in patches 1 and 2, and highest energy density was applied in patches 3 and 4. A gas flow velocity of 6.7 m/s was found to be adequate to remove process byproducts under the conditions tested (to be discussed in more detail further in this section), and is representative of adequate gas flow conditions, as shown in the left column of Figure 4. The right column of Figure 4 shows meltpools generated with no gas flow.

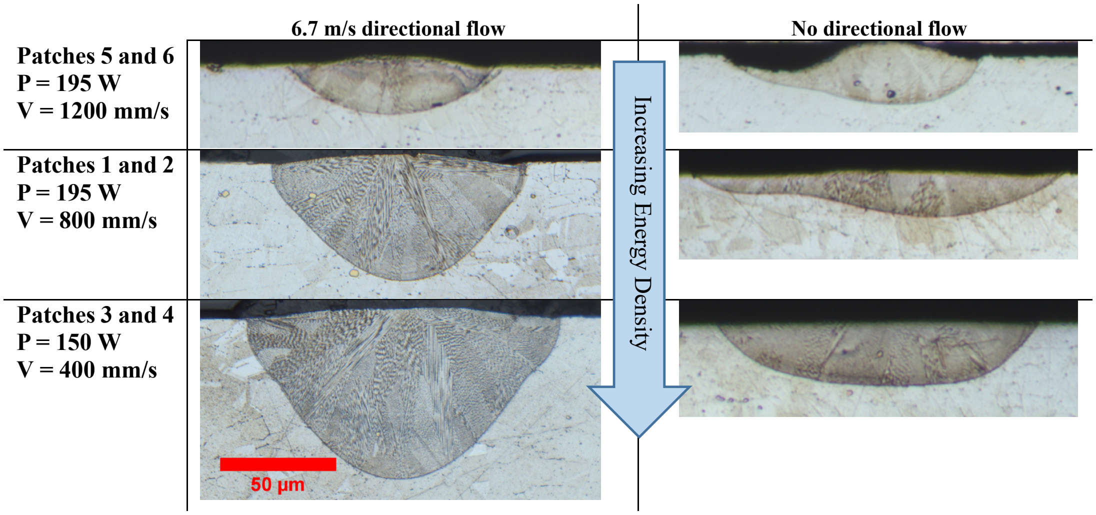


Figure 4: Cross-section views of melt tracks generated with varying energy density and gas flow speed. All tests were with IN625 in an argon environment. All images are same scale.

Starting with the meltpools generated with 6.7 m/s gas flow, the lowest energy density used in patches 5 and 6 forms meltpools in conduction mode with a low AR. With the increased energy density in patches 1 and 2 and 6.7 m/s gas flow, the meltpool width increases slightly and the depth increases significantly, with the AR increasing proportionally. The meltpool mode of patches 1 and 2 is in transition mode. With the highest energy density tested in patches 3 and 4, the meltpool width and depth both increase to the maximum observed. The track generated in patches 5 and 6 are toward keyhole mode than patches 3 and 4, but still in transition mode. All tracks generated with 6.7 m/s gas flow were highly symmetric about their centerline.

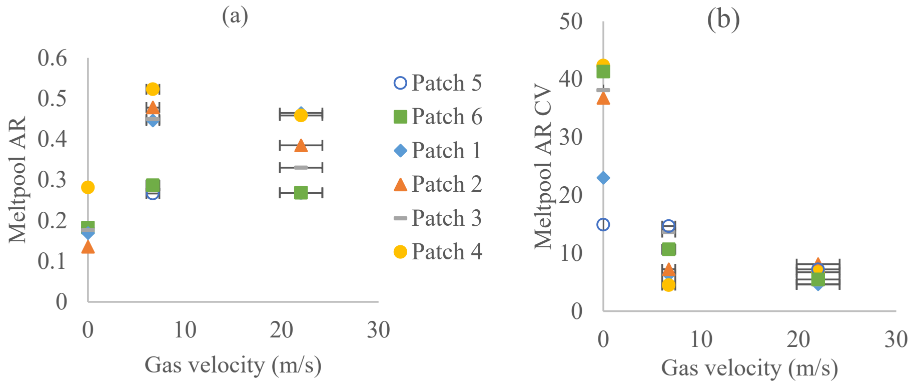


Figure 5: (a) meltpool AR (a) and meltpool AR CV (b) as a function of gas flow speed. Odd numbered patches were scanned nominally perpendicular to the gas flow with hatching direction nominally toward the flow and even numbered patches were scanned nominally toward the gas flow with hatching direction was nominally perpendicular the flow. The lowest energy density was applied to patches 5 and 6, medium energy density was applied to patches 1 and 2, and highest energy density was applied to patches 3 and 4.

The detriment of no directional gas flow is evident in the tests with no directional gas flow, shown in the right column of Figure 4. The melt tracks are consistently asymmetric about their centerline, while also being wider and shallower than those with adequate gas flow. The meltpool depth generally increased with increasing energy density but was highly irregular. These results are consistent with beam scattering, attenuation, and lensing may widen the beam and form a shallower and wider meltpool [11,13].

Figure 5a shows the measured meltpool AR as a function of gas flow speed. With no gas flow, all patches exhibit low AR, generated by shallow and wide meltpools. Because the low energy density tracks (patches 5 and 6) are in conduction mode, they generate little to no vapor, and the low meltpool AR exhibited with no gas flow may be a product of residual suspended condensate from the previous higher energy density scans. The higher energy density (patches 1 through 4) all exhibit their highest AR at 6.7 m/s. Then, AR decreases slightly for the higher energy density patches from 6.7 m/s to 22 m/s. This decrease in AR is associated with reduced laser delivery and may be a product of excessive gas flow causing rapid condensation directly above the meltpool, which causes slight beam attenuation. This inference is consistent with the very small change observed in the conduction mode patches (patches 5 and 6) from 6.7 m/s to 22 m/s, because there is little or no vapor generated, and therefore, little or no possibility of rapid condensation above the meltpool.

As shown in Figure 5b, increased gas flow rate is associated with decreasing meltpool AR coefficients of variation (CV). With no directional gas flow, patches 4 and 6 exhibited a CV of more than 40, indicating a highly inconsistent process. From 6.7 m/s to 22 m/s, some patches increased slightly in variability and
some decreased slightly, with no clear trend. But, as a whole, the scans with 22 m/s showed less variability than the scans at 6.7 m/s, which appears to be due to the effective (although excessively high velocity) removal of process byproducts from the beam path. Neither the meltpool AR nor AR CV showed a clear dependence on gas flow direction, with both varying stochastically from patch to patch with x-direction scans compared with y-direction scans, as shown in the appendix in Figure 13 and Figure 14.

In summary, the meltpool AR showed only slight changes with a gas flow velocity above 6.7 m/s and the meltpool AR CV fell below 15 % above 6.7 m/s. Therefore, it was a gas flow velocity of 6.7 m/s was said to be “adequate” gas flow to facilitate a consistent process without adverse effects of byproduct accumulation.

3.2 Effects of Gas Flow Speed on Meltpool Signature and Hot Vapor Plume

Figure 6a and b show the synchronized view of the inline meltpool monitoring camera and an oblique view camera at 2,000 fps with adequate gas flow and no gas flow, respectively. As it can be seen that with adequate gas flow (22 m/s), the inline hotspot remains a consistent size and shape from frame to frame. Similarly, with adequate gas flow, the hotspot in the oblique view remains a consistent size and shape. The hot vapor jet can be seen directly above the oblique view hot spot, and although the jet size, shape, and direction vary slightly, its general profile remains quite consistent throughout the sequence.

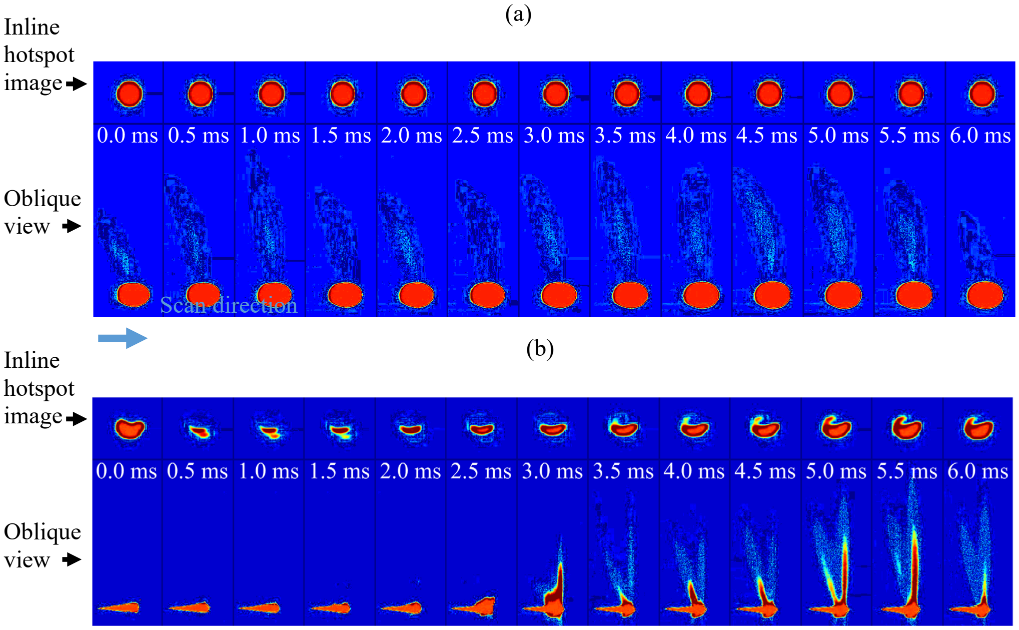


Figure 6: inline hotspot image synchronized with an oblique view of meltpool and vapor plume with (a) 22 m/s directional gas flow and (b) no directional gas flow. Both tracks were melted in an argon environment on IN625 with a 1070 nm fiber laser, a laser spot D4σ of 100 µm, a scan speed of 400 mm/s, and laser power 150 W. The inline and oblique view images were taken at nominally 850 nm wavelength. Images are gamma-adjusted to enhance contrast, so intensity is not directly indicative of temperature.

In contrast with adequate gas flow, it can be seen in Figure 6b that the inline hotspot image is changing significantly in size and shape along the length of the track. The hotspot in the oblique view changes in size and shape slightly, but the vapor jet changes significantly along the length of the track. From 0.0 ms to 2.5 ms, no vapor jet is visible, suggesting that the beam energy density is diminished so much that vapor generation is not occurring during that period, indicating that the beam is scattered and attenuated by lingering condensate generated by previous tracks. After 2.5 ms, the vapor jet re-forms with rapidly changing size and shape until its endpoint at 6.0 ms. This obstruction of beam delivery is consistent with the findings of Shcheglov et al. [10] when gas flow inadequately removed process byproducts from the beam path in laser welding. In comparing Figure 6a and b, it is evident that a large and consistent inline hotspot size is associated with a consistent vapor jet from the meltpool.

Figure 7a shows an optical image of a set of melt tracks (all 6 patches) generated with 22 m/s directional gas flow. In Figure 7a, the tracks are visibly consistent in width and texture along the length of each track and repeatable from track to track. The chevron patterns are relatively evenly spaced and the terminal meltpool at the end of each track have consistent sizes for each set of process parameters. Figure 7b shows a contour plot of the inline image pixel count (above 80 digital levels) at each location along scan path of the track shown in Figure 7a. The number of pixels is proportional to the area of the laser-metal interaction. The hotspot area is largest in patches 1 through 4, at about 600 pixels to more than 800 pixels. The large hotspot area is generated by the higher laser energy density applied in those patches. The smallest hot spot is generated in patches 5 and 6, in the range of about 400 pixels because of the lower energy density applied to those patches. The hotspot size is quite consistent and repeatable along the length of each line and from line to line.

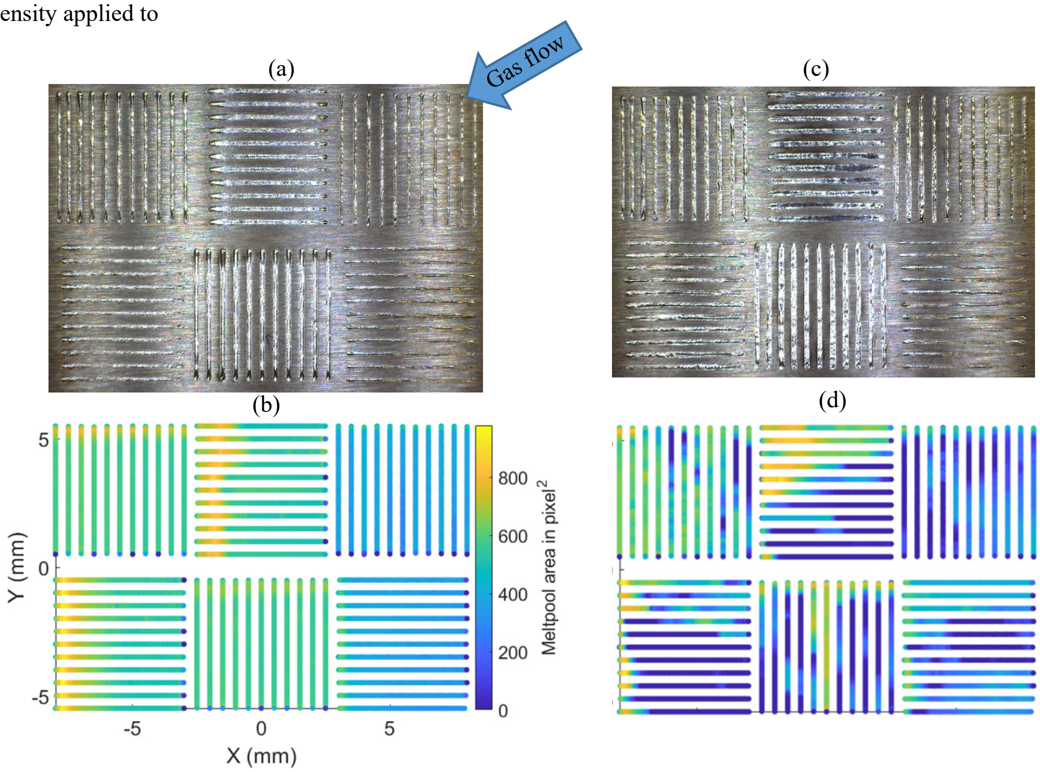


Figure 1: (a) surface image of tracks melted with 22 m/s directional gas flow, (b) contour plot of inline image hotspot pixel count at each location along tracks of (a), (c) surface image of tracks melted with no directional gas flow, (d) contour plot of inline image hotspot pixel count at each location along tracks of (c)

In contrast to Figure 7a, Figure 7c shows that the melt tracks are visibly inconsistent with no directional gas flow applied to the process. The melt tracks are unstable, unrepeatable, and highly variable in width and texture. The hotspot contour map in Figure 7d shows high variability in the hotspot size, ranging from nearly 0 to 800 pixels sporadically along the length of the tracks and from track to track. Areas of small hotspot size appear to correlate with visible widenings in the melt tracks. The only track with low variability is the first track scanned (furthest left in patch 1) because of the lack of accumulated suspended condensate in the beam path.

Figure 8a shows the relationship between hotspot area and gas flow speed. From 0 m/s to 6.7 m/s all patches show an increasing trend in hotspot area with gas flow speed, with slightly more scatter occurring at 2.2 m/s compared with 0 m/s. From 6.7 m/s to 22 m/s, each patch hotspot size decreases slightly. This is likely due to rapid condensation of vapor directly above the meltpool caused by the excess gas flow, as was discussed previously. This condensate directly above the meltpool may attenuate the beam and reduce the beam energy density delivered to the process. The condensate directly above the meltpool may also attenuate the light emitted from the process and detected by the inline imager. No clear trend was associated with hot spot size and scan direction relative to the gas flow direction, as is shown in Appendix Figure 12.

Deisenroth, David; Neira, Jorge; Weaver, Jordan; Yeung, Ho. "Effects of Shield Gas Flow on Meltpool Variability and Signature in Scanned Laser Melting." Presented at ASME 2020 15th International Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

Figure 8b shows that from 0 m/s to 6.7 m/s all patches show a monotonic decrease in hotspot CV, with the exception of patch 5 at 2.2 m/s. After 6.7 m/s, all patches show a slight increase in the hotspot variability. The slight increase in hotspot variability at 22 m/s may be due to increased emission and emission variability from the rapidly condensing

vapor jet directly above the meltpool, essentially generating a "flickering" effect and an apparent increase in hotspot variability interpreted by the inline imager. The trend observed in hot spot

size variability with meltpool AR variability as a function of scan direction will be discussed in the following section.

3.3 Relationship Between Hotspot Area and Meltpool Aspect Ratio

One of the intentions of the current investigation was to better understand the relationship between inline image hotspot size and meltpool AR, and the results are shown in Figure 9. With no gas flow, there is a significant amount of scatter in the average hotspot size produced in each patch; the hotspot size is also generally smaller with no gas flow than with ≥6.7 m/s in each patch. With gas flow ≥6.7 m/s there is an increasing trend with hotspot area and meltpool AR in each path. In the low energy density patches in conduction mode (patches 5 and 6), that generate an AR of about 0.3, a hotspot size of about 0.25 mm^2 was generated. In the higher energy density patches (patches 1 through 4) with AR greater than 0.3, the hotspot size increased to about 0.04 mm^2. It can be observed that in transition mode (AR > 0.3), HS area increases only slightly with a significant increase in meltpool AR. This lack of sensitivity of hotspot area to meltpool AR indicates that hotspot area is a diminishingly

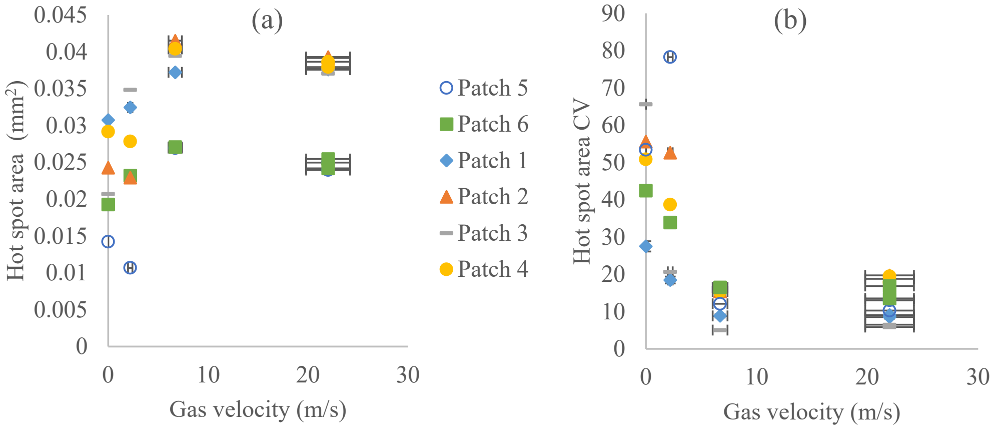


Figure 8: (a) hotspot area and (b) hotspot area cv as function of directional shield gas flow rate

useful indicator as the process transitions toward keyhole mode. Under the conditions tested, the HS area could be used to discern between conduction mode and transition mode processes, but may not be a useful process signature for processes with higher energy densities than early transition mode.

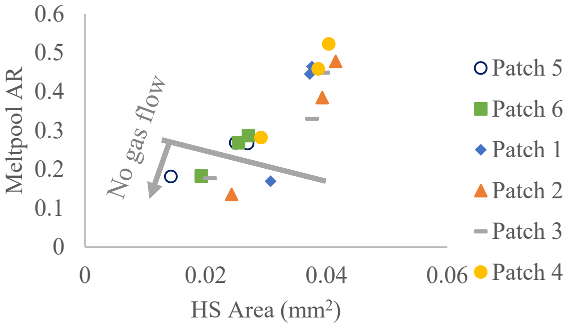


Figure 9: hotspot area as a function of meltpool AR

Hotspot area CV as a function of meltpool AR coefficient of variability was found to be a stronger indicator of the extreme process variability associated with inadequate gas flow than average hotspot size was, as shown in Figure 10. Meltpool AR coefficients of variation greater than 14.8 were associated with no gas flow. With no gas flow, the hotspot area CV increased by a factor of about two when compared to the hot spot CV generated with adequate gas flow of ≥6.7 m/s, producing a clear indication of meltpool AR variability. It can be seen, though, that with gas flow ≥6.7 m/s, the hotspot area coefficient of variability splits into two distinct groups: x-scan direction and y-scan direction. With gas flow ≥6.7 m/s and comparable meltpool AR coefficients of variation, the hotspot area coefficients of variation show a significant bias toward higher values when scanning in the x-direction, which is toward the gas flow. With scan direction relative to gas flow direction being the only variable changed, it is likely that a plume effect is the cause of the increased "flicker" in the inline hotspot size.

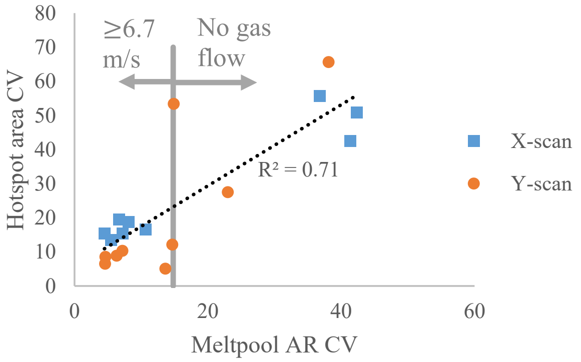


Figure 10: hotspot area cv as a function of meltpool ar variability. X-scans are nominally toward the gas flow, y-scans are nominally perpendicular to the gas flow

The inline hotspot area CV generated with patches scanned in the x-direction compared with the y-direction are shown in Figure 11. As would be expected, there is a significant amount of scatter with no discernable pattern in the region with no gas flow, which exhibits coefficients of variation greater than 30. But, a quite clear trend emerges in the region with gas flow ≥6.7 m/s, in which the amount of bias toward higher hotspot variability increases with increasing process energy density.

With patches 5 and 6 in conduction mode, little or no vapor is generated, and the bias toward higher variability in the x-direction is small with gas flow ≥6.7 m/s. The bias increases with higher process energy density. At the highest process energy density, scans in the x-direction exhibit 3 to 4 times higher hotspot coefficients of variation with gas flow ≥6.7 m/s. This increasing variability in hotspot size, therefore, appears to be a function of laser interaction with the hot vapor jet differently when scanning nominally toward the gas flow compared with nominally perpendicular to the gas flow. It seems particularly likely that the increased variability in hotspot size with scan direction is due to the hot vapor jet considering that no clear scan-direction bias was found in meltpool AR, meltpool AR CV, or hotspot area. The cause of hotspot area CV changes with scan direction are currently unknown but may be related to increased vapor jet velocity relative to the gas flow and/or changes in vapor jet incline angle with gas flow direction.

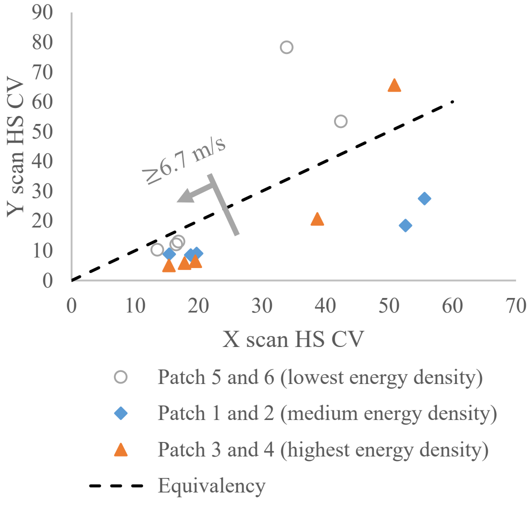


Figure 11: inline hotspot CV at three power densities with x and y scan directions

4. Conclusion

A preliminary study on the effects of gas flow speed and direction on meltpool aspect ratio and hotspot size was performed under process parameters similar to those used in LPBF in IN625. It was found that no gas flow was associated with asymmetric, shallow, and wide melt tracks that may cause lack of fusion defects. The melt tracks became consistent in width, depth, and shape with adequate gas flow. Excessive gas flow was associated with higher consistency in meltpool aspect ratio, but shallower meltpools.

For process monitoring, it was found that, under the conditions tested, inline image hotspot size increased with energy densities that formed meltpools ranging from conduction mode to transition mode. Average hotspot area was, though, a relatively insensitive indicator of both meltpool aspect ratio and high process variability, especially as the energy density increased in the transition mode toward keyhole mode. Hotspot area coefficient of variation was found to be a stronger indicator of meltpool aspect ratio variability, and therefore an indicator of inadequate gas flow. Use of hotspot area coefficient of variation as an indicator of meltpool aspect ratio variability suffered from a bias toward higher variability when the meltpool was scanned nominally toward the gas flow. The bias that caused higher “flicker” in the meltpool image hotspot area was likely due to a change in the laser interaction with the hot vapor jet with different gas flow directions. Therefore, it is concluded that gas flow speed and direction each have important ramifications for both laser delivery and process monitoring.

Future work will be in developing gas flow profiles (especially increased height of the velocity profile) that are compatible with LPBF powders, as the velocities used in this study would likely disrupt the powder bed. This study has improved the understanding of the relationship between meltpools and meltpool signature, but more work is needed to develop robust process monitoring. Meltpool image intensity and intensity variability will be investigated in future work.

ACKNOWLEDGEMENTS

The authors wish to thank Brandon Lane of the NIST Engineering Laboratory for his help with uncertainty estimates and Sergey Mekhontsev of the NIST Physical Measurement Laboratory for his aid in setup of the plume visualization experiments.

REFERENCES

[1] Coeck, S., Bisht, M., Plas, J., and Verbist, F., 2019, “Prediction of Lack of Fusion Porosity in Selective Laser Melting Based on Melt Pool Monitoring Data,” Additive Manufacturing, 25, pp. 347–356.

[2] Ferrar, B., Mullen, L., Jones, E., Stamp, R., and Sutcliffe, C. J., 2012, “Gas Flow Effects on Selective Laser Melting (SLM) Manufacturing Performance,” Journal of Materials Processing Technology, 212(2), pp. 355–364.

[3] Bean, G. E., Witkin, D. B., McLouth, T. D., and Zaldivar, R. J., 2018, “The Effect of Laser Focus and Process Parameters on Microstructure and Mechanical Properties of SLM Inconel 718,” Laser 3D Manufacturing V, International Society for Optics and Photonics, p. 105230Y.

[4] Kong, C.-J., Tuck, C. J., Ashcroft, I. A., Wildman, R. D., and Hague, R., 2011, “High Density Ti6Al4V via SLM Processing: Microstructure and Mechanical Properties,” International Solid Freeform Fabrication Symposium, pp. 475–483.

[5] 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. Advanced Manufacturing Technology, 95(1–4), pp. 527–550.

[6] Elmer, J. W., Vaja, J., Carlton, H. D., and Pong, R., 2015, “The Effect of Ar and N2 Shielding Gas on Laser Weld Porosity in Steel, Stainless Steels, and Nickel,” Weld J., 94(10), pp. 313s–325s.

[7] Ahn, J., He, E., Chen, L., Dear, J., and Davies, C., 2017, “The Effect of Ar and He Shielding Gas on Fibre Laser Weld Shape and Microstructure in AA 2024-T3,” J. Manufacturing Processes, 29, pp. 62–73.

[8] 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), p. 4085.

[9] Shcheglov, P., 2012, “Study of Vapour-Plasma Plume during High Power Fiber Laser Beam Influence on Metals,” Ph.D. thesis, BAM Federal Institute for Materials Research and Testing.

[10] Shcheglov, P. Y., Gumenyuk, A. V., Gornushkin, I. B., Rethmeier, M., and Petrovskiy, V. N., 2012, “Vapor–Plasma Plume Investigation during High-Power Fiber Laser Welding,” Laser Physics, 23(1), p. 016001.

[11] Katayama, S., Kawahito, Y., and Mizutani, M., 2010, “Elucidation of Laser Welding Phenomena and Factors Affecting Weld Penetration and Welding Defects,” Physics Procedia, 5, pp. 9–17.

[12] Zhirnov, I., Kotoban, D. V., and Gusarov, A. V., 2018, “Evaporation-Induced Gas-Phase Flows at Selective Laser Melting,” Applied Physics A, 124(2), p. 157.

[13] Ladewig, A., Schlick, G., Fisser, M., Schulze, V., and Glatzel, U., 2016, “Influence of the Shielding Gas Flow on the Removal of Process By-Products in the Selective Laser Melting Process,” Additive Manufacturing, 10, pp. 1–9.

[14] Anwar, A. B., and Pham, Q., 2016, “Effect of Inert Gas Flow Velocity and Unidirectional Scanning on the Formation and Accumulation of Spattered Powder during Selective Laser Melting,” 2nd Intl. Conf. on Progress in Additive Manufacturing, Singapore, pp. 531 – 536.

[15] Aboulkhair, N. T., Everitt, N. M., Ashcroft, I., and Tuck, C., 2014, “Reducing Porosity in AlSi10Mg Parts Processed by Selective Laser Melting,” Additive Manufacturing, 1, pp. 77–86.

[16] Weaver, J. S., Kreitman, M., Heigel, J. C., and Donmez, M. A., 2019, “Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation,” TMS 2019 148th Annual Meeting & Exhibition Supplemental Proceedings, Springer, pp. 269–278.

[17] Keller, T., Lindwall, G., Ghosh, S., Ma, L., Lane, B. M., Zhang, F., Kattner, U. R., Lass, E. A., Heigel, J. C., and Idell, Y., 2017, “Application of Finite Element, Phase-Field, and CALPHAD-Based Methods to Additive Manufacturing of Ni-Based Superalloys,” Acta Materialia, 139, pp. 244–253.

[18] 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.

[19] Yadroitsau, I., 2009, Selective Laser Melting: Direct Manufacturing of 3D-Objects by Selective Laser Melting of Metal Powders, Lambert Academic Publishing.

[20] Trapp, J., Rubenchik, A. M., Guss, G., and Matthews, M. J., 2017, “In Situ Absorptivity Measurements of Metallic Powders during Laser Powder-Bed Fusion Additive Manufacturing,” Applied Materials Today, 9, pp. 341–349.

[21] Ye, J., Khairallah, S. A., Rubenchik, A. M., Crumb, M. F., Guss, G., Belak, J., and Matthews, M. J., 2019, “Energy Coupling Mechanisms and Scaling Behavior Associated with Laser Powder Bed Fusion Additive Manufacturing,” Advanced Engineering Materials, p. 1900185.

[22] Simonds, B. J., Sowards, J., Hadler, J., Pfeif, E., Wilthan, B., Tanner, J., Harris, C., Williams, P., and Lehman, J., 2018, “Time-Resolved Absorptance and Melt Pool Dynamics during Intense Laser Irradiation of a Metal,” Phys. Rev. Applied, 10(4), p. 044061.

[23] Spears, T. G., and Gold, S. A., 2016, “In-Process Sensing in Selective Laser Melting (SLM) Additive Manufacturing,” Integrating Materials and Manufacturing Innovation, 5(1), pp. 16–40.

[24] Ye, D., Zhu, K., Fuh, J. Y. H., Zhang, Y., and Soon, H. G., 2019, “The Investigation of Plume and Spatter Signatures on Melted States in Selective Laser Melting,” Optics & Laser Technology, 111, pp. 395–406.

[25] Tenner, F., Brock, C., Klämpfl, F., and Schmidt, M., 2015, “Analysis of the Correlation between Plasma Plume and Keyhole Behavior in Laser Metal Welding for the Modeling of the Keyhole Geometry,” Optics and Lasers in Engineering, 64, pp. 32–41.

[26] Lane, B., Mekhontsev, S., Grantham, S., Vlasea, M., Whiting, J., Yeung, H., Fox, J., Zarobila, C., Neira, J., and McGlauflin, M., 2016, “Design, Developments, and Results from the NIST Additive Manufacturing Metrology Testbed (AMMT),” Solid Freeform Fabrication Symposium, Austin, TX, pp. 1145–1160.

[27] Levine, L., Lane, B., Heigel, J., Migler, K., Stoudt, M., Phan, T., Ricker, R., Strantza, M., Hill, M., Zhang, F., Seppala, J., Garboczi, E., Bain, E., Cole, D., Allen, A., Fox, J., and Campbell, C., 2020, “Outcomes and Conclusions from the 2018 AM-Bench Measurements, Challenge Problems, Modeling Submissions, and Conference,” Integrating Materials and Manufacturing Innovation.

[28] ISO, 1997, “Geometrical Product Specifications (GPS)–Surface Texture: Profile Method–Terms, Definitions and Surface Texture Parameters,” 4287.

[29] Zhirnov, I., 2019, “Dynamic Measurement of Laser Beam Quality for Selective Laser Melting,” Presented at Solid Freeform Fabrication Symposium, Austin.

Appendix

As shown in Figure 12, there is no clear gas flow direction bias in hotspot area. As shown in Figure 13, no trend was observed due to scan direction in meltpool AR. As shown in Figure 14, no clear bias is shown in meltpool AR CV due to scan direction.

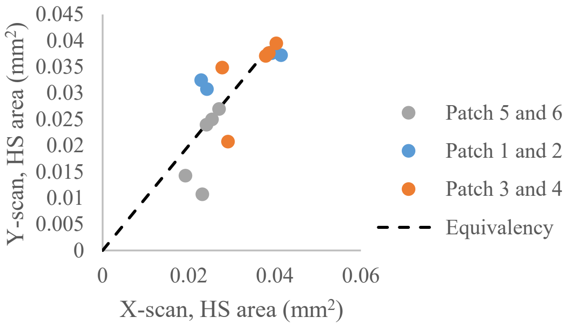


Figure 12: hot spot size bias with scan direction

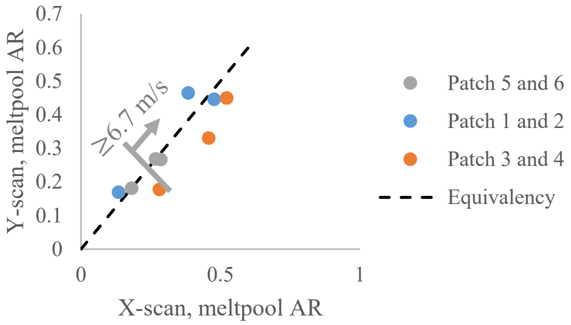


Figure 13: meltpool aspect ratio direction dependence

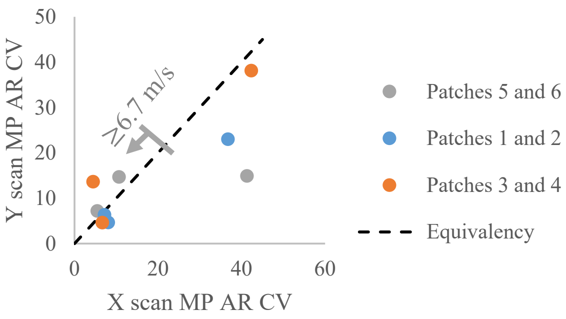


Figure 14: meltpool aspect ratio cv direction dependence

Deisenroth, David; Neira, Jorge; Weaver, Jordan; Yeung, Ho. "Effects of Shield Gas Flow on Meltpool Variability and Signature in Scanned Laser Melting." Presented at ASME 2020 15th International Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

Cryogenic Calibration of a Quantum-based Radio Frequency Source

A. S. Boaventura[#&1], J. A. Brevik[#], D. F. Williams[#], A. E. Fox[#], M. C. Beltran[#], P. F. Hopkins[#], P. D. Dresselhaus[#], S. P. Benz[#]
[#]National Institute of Standards and Technology, Boulder, CO 80305 USA
[&]Department of Physics, University of Colorado Boulder,
[1]aliriodejesus.soaresboaventura@nist.gov

***Abstract*— We report on the calibration of quantum-based radio frequency waveforms generated by a Josephson arbitrary waveform synthesizer system. We measure these waveforms using a vector network analyzer and calibrate them at 4 K using a custom-designed cryogenic on-wafer multi-line thru-reflect-line calibration kit and a two-tier calibration procedure. The signals tested in this work can be used as reference signals to calibrate measurement instruments with potential benefits in terms of accuracy and flexibility compared to the conventional methods.**

***Keywords* — superconductive circuits, cryogenic microwave measurements, on-chip calibration, Josephson arbitrary waveform synthesizer, quantum-based voltage standards.**

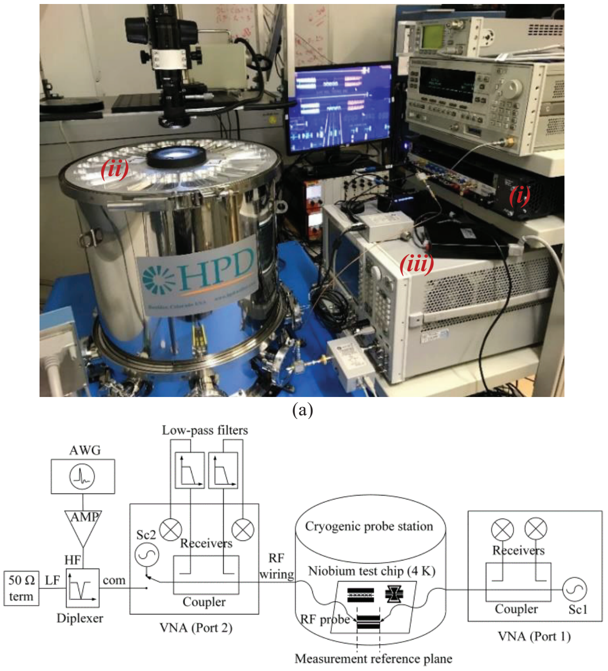
I. Introduction

The Josephson arbitrary waveform synthesizer (JAWS) system has been used to generate voltage standards for metrology at audio frequencies [1]. We are currently extending the JAWS capability to the microwave frequency range for use in wireless communications metrology. While signal characterization and calibration are not required at audio frequencies, they are crucial if the JAWS system is to be used for metrology at microwave frequencies.

Although the RF signals generated by the JAWS system present quantum-based accuracy across the on-chip 4 K Josephson junction (JJ) circuit, the accuracy is lost off chip as the signals are conducted to room-temperature through long RF cables that introduce losses and phase offsets. Thus, traceable RF calibration is imperative to accurately translate the measurements made at room-temperature to the reference plane of the 4 K JJ circuit and retrieve the quantum on-chip accuracy.

The JAWS system exploits the voltage pulse quantization of cryo-cooled 4 K JJs to generate quantum-based waveforms with calculable amplitudes that are related to fundamental constants [1]-[3]. Driving an array of N JJs with a single current pulse yields a voltage pulse across the array with a time-averaged area equal to NΦ_0, where Φ_0 is the magnetic flux quantum. To synthesize an arbitrary RF waveform, the JJ array is driven with a train of high-speed pulses whose separation is modulated with a delta-sigma algorithm to encode the desired waveform [4]. Typically, the driving pulse pattern is created by a room-temperature arbitrary waveform generator (AWG), and the pulse pattern produced by the array is low-pass filtered to get the quantized RF waveform.

The JAWS system has been used to synthesize waveforms up to 1 GHz [5], but those waveforms were not calibrated. In this work, we synthesize quantum-based RF waveforms using the JAWS system and calibrate them at 4 K using a vector network analyzer (VNA). For that, we add absolute amplitude and phase correction [6] to a scattering-parameter calibration [7] to provide a full wave-parameter calibration of the JAWS waveforms at the reference plane of the 4 K JJ circuit.



(a)

(b)

Fig. 1 a) Photograph of the cryogenic measurement setup including the room-temperature AWG that generates the driving patterns *(i)*, 4 K cryogenic probe station with test chip that contains the JJ circuit and calibration standards *(ii)*, and VNA apparatus to measure the JAWS signals *(iii)*. b) Simplified diagram of the measurement setup shown above. Low-pass filters are used at the input of the VNA receivers to improve the dynamic range of our measurements.

The quantum-based RF waveforms we test in this work can be used, for example, as reference signals to calibrate RF measurement instruments with potential benefits in terms of accuracy and flexibility compared to the conventional methods.

II. Measurement and Calibration Procedures

A. *Measurement setup and quantum signal generation*

The cryogenic measurement setup used in this work is shown in Fig.1. We use a high-speed arbitrary waveform generator (AWG) to create a delta-sigma-modulated current pulse pattern. This pulse pattern encodes the RF waveform to be quantized by the JJ circuit operating in the cryogenic probe station at 4 K [4].

We measure the JAWS signal in the frequency domain using a VNA (PNA-X[1]) that contacts the 4 K JAWS chip via cryogenic RF cables and movable RF probes installed on the 4 K stage of the cryogenic probe station. Internal switches allow us to route the VNA test ports to either the VNA internal

[1] We specify equipment models only to better explain the experiments. NIST does not endorse commercial products. Other products may perform as well or better.

Soares Boaventura, Alirio De Jesus; Brevik, Justus; Williams, Dylan; Fox, Anna; Castellanos Beltran, Manuel; Hopkins, Pete; Dresselhaus, Paul; Benz, Samuel P. "Cryogenic Calibration of a Quantum-based Radio Frequency Source." Presented at 95th ARFTG Microwave Measurement Conference, Los Angeles, CA, US. August 04, 2020 - August 06, 2020.

continuous wave (CW) sources (during the VNA calibration phase) or the external AWG driving pulse source (during the JAWS measurement phase).

To provide enough amplitude to properly drive the JJ circuit, we amplify the pulse driving pattern signal generated by the AWG using a broadband RF amplifier. We use a diplexer to isolate the JAWS output signal from the input pulse driving signal. In addition, the diplexer's low-frequency port is terminated with a 50 Ω load to avoid reflection of the JAWS signal towards the 4 K JJ circuit. We use a 16 GHz sine wave signal generated by an RF signal generator (not shown in Fig. 1b) as the primary phase reference for our measurements. This 16 GHz signal drives the AWG which in turn clocks the VNA with its 10 MHz output clock signal and drives the VNA comb generators with a 1 MHz square-wave signal that sets our measurement frequency grid.

We measured the frequency response of the JAWS output signal by measuring the power and phase at the fundamental of a single-tone signal that was synthesized in steps of 1 MHz from 10 MHz to 1 GHz. The 991 single-tone bipolar waveforms were each encoded into bias pulse patterns with a minimum of 10,000 waveform periods using a second-order, three-level [-1; 0; +1], bandpass delta-sigma modulator at a 64 gigapulse-per-second sample rate.

It is crucial that the driving pulse pattern has minimal power at the synthesis frequency. To reduce the signal level at the fundamental in the driving pulse pattern, the three-level codes were transformed into five-level codes so that each bias pulse had bracketing half-amplitude pulses with inverted amplitude and each pulse block had effectively zero average amplitude [3].

Each of the 991 pulse patterns was then sequentially programmed into the AWG, the VNA measurement frequency was set to the corresponding synthesis frequency, and the waveforms incident at and reflected from the JJ circuit were measured. For this experiment, quantum-locked operation of the JAWS system could be achieved across the entire synthesis frequency range only by reducing the amplitude of the waveforms to 2.5 % of the maximum pulse density of the input driving pulse train [4].

B. *Measurement calibration procedure*

The three-step procedure to calibrate the VNA measurements of the JAWS waveforms is illustrated in Fig. 2. This procedure allows us to operate our VNA with an external source and evaluate the calibrated forward propagating wave (AWG pulse driving signal) and backward propagating wave (JAWS output signal) at the 4 K on-chip reference plane.

1. First-tier calibration: we first operated the VNA in its default configuration (Fig. 2a), with the test ports routed to the internal CW sources, and performed a scattering-parameter, absolute amplitude, and absolute phase calibration at the coaxial reference plane in steps of 1 MHz from 10 MHz to 1 GHz [6].

The scattering-parameter calibration was performed using a 2.4 mm short-open-load-thru (SOLT) calibration kit, the phase was calibrated using a frequency comb generator that is traceable to the NIST electro-optic sampling system, and the amplitude calibration was done with a power sensor that is traceable to the NIST calorimetric power reference. By performing the first-tier calibration, we effectively set the calibration reference plane at the coaxial plane of the VNA test ports.

2. Second-tier calibration: in the second calibration step (Fig. 2b), we connected the VNA (still in its default configuration) to the cryogenic probe station and performed a 4 K scattering-parameter calibration at the reference plane of the JJ circuit [7]. For that, we used a custom-designed cryogenic multi-line TRL calibration kit fabricated on the same chip as the JJ circuit. This kit comprises two coplanar waveguide (CPW) reflect standards (open and short), six CPW lines with lengths ranging from 70 μm to 9 mm, and a 0-length CPW thru line that sets the reference plane of our TRL calibration exactly at the terminals of the JJ circuit. Additionally, to allow simple on-wafer SOLT calibrations, we included 50 Ω load standards in the kit. By performing the second-tier calibration, we move the reference plane from the VNA coaxial plane to the on-chip reference plane of the JJ circuit. Our calibration procedure has a 50 GHz bandwidth capability, which allows to support future frequency scaleup of the JAWS system.

3. Measurement: for the JAWS measurements, we routed the switching circuitry of the VNA test port 2 to the AWG connected to the back-panel of the VNA (see Fig. 2c), and measured the JAWS signal that was generated as described in section II.A. Note that changing the VNA source configuration does not affect the calibration performed in steps 1 and 2, because the full wave-parameter calibration matrix is source-independent. To correct the measured data, we used the NIST Microwave Uncertainty Framework (MUF) software [8].

III. Measurement Results

A. *Control experiment*

In order to help verify our calibration, we used the setup depicted in Fig. 3 to mimic the JAWS measurements. For that, we created a swept-sine signal from 10 MHz to 1 GHz in steps of 1 MHz using an RF signal generator, fed this signal into the cryogenic probe station port 1 and measured it at the VNA test port 2 via the on-chip thru standard.

We calibrated the measured data of this swept-sine at the TRL reference plane like we did for the JAWS waveforms and de-embedded the cable of port 1 to reference the measurements to the coaxial connector of the signal generator (see Fig. 3). The measured results presented in Table 1 agree with the nominal power levels generated by the signal generator within the manufacturer margins. Although this measurement setup is simpler than the JAWS setup as there are no broadband pulse signals involved and no linearity issues, it gives a good indication that our amplitude calibration is correct.

Soares Boaventura, Alirio De Jesus; Brevik, Justus; Williams, Dylan; Fox, Anna; Castellanos Beltran, Manuel; Hopkins, Pete; Dresselhaus, Paul; Benz, Samuel P. "Cryogenic Calibration of a Quantum-based Radio Frequency Source." Presented at 95th ARFTG Microwave Measurement Conference, Los Angeles, CA, US. August 04, 2020 - August 06, 2020.

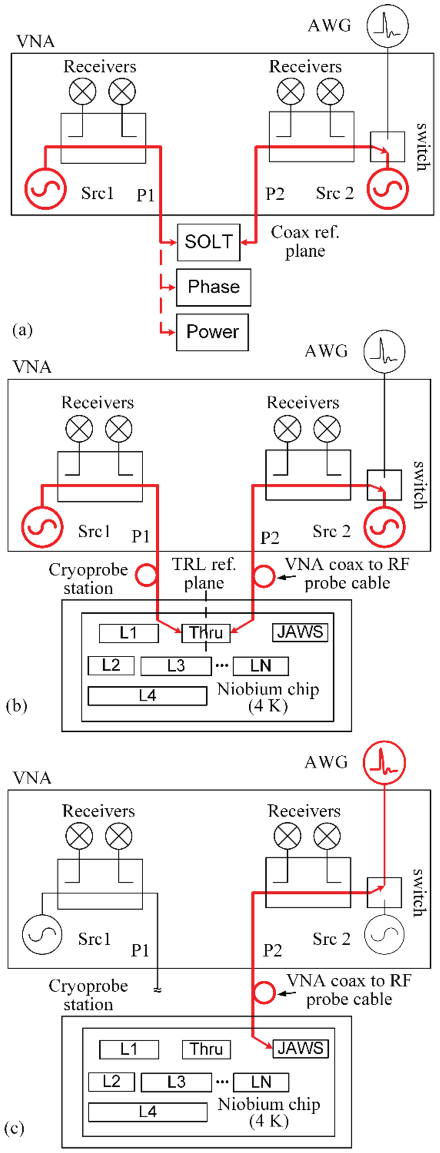


Fig. 2a) First-tier VNA calibration using SOLT, power and phase standards. b) Second-tier VNA calibration using broadband on-wafer multiline TRL kit. c) JAWS measurements, with AWG driving signal passed through the VNA.

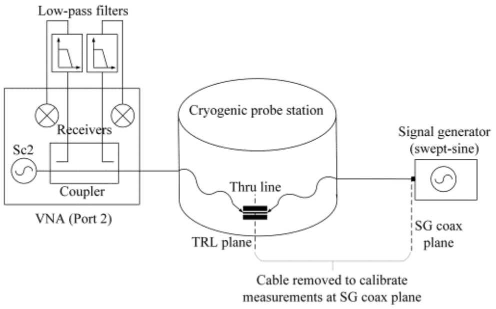


Fig. 3 Simple control experiment used to verify the calibration.

Table 1 Calibrated signal generator amplitude. The average and standard deviation are taken over the 1 GHz bandwidth measurements.

Nominal output power level (dBm)	Calibrated mean	Standard deviation
0 ± 0.5	– 0.07	0.05
– 30 ± 0.7	– 30.07	0.05

B. *Optimizing the VNA dynamic range*

To generate a signal amplitude high enough to properly drive the JJ circuit, we amplified the pulse pattern signal using a broadband RF amplifier prior to sending it through the VNA test set. The grey curve in Fig. 4a shows the 50 GHz spectrum measured by the VNA for a delta-sigma pulse pattern used in our experiments. This signal has frequency components that are much larger than the low-frequency JAWS output signal of interest.

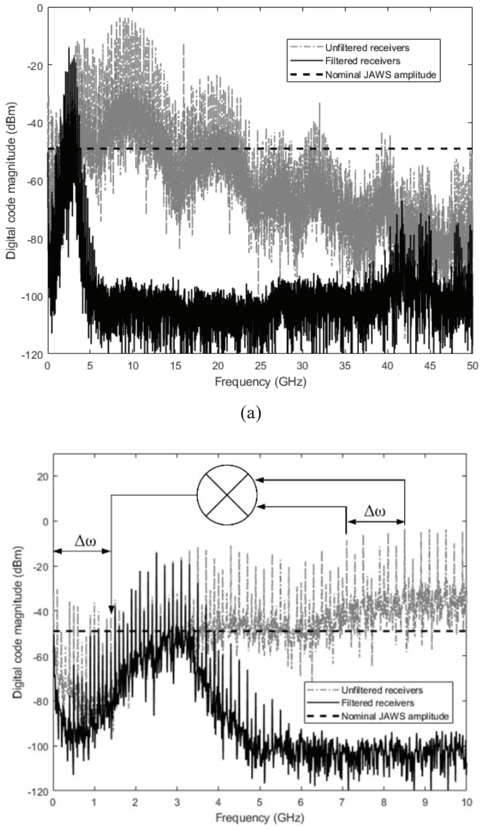
Due to non-linear intermodulation distortion in the VNA receivers, large high-frequency components of the driving pulse pattern (which effectively is a broadband multi-tone signal) can be down-converted and lead to erroneous measurement and calibration of the low-frequency JAWS signal of interest. To overcome this problem, which is simplistically illustrated in the inset in Fig. 4b and detailed in [9], we used low-pass filters with 1 GHz cut-off frequency at the input of the VNA port 2 receivers (see filtered signal in Fig. 4a). This produced measurement results more consistent with the expected JAWS frequency response predicted in [4] (dashed lines in Fig. 4 and Fig. 5).

C. *Calibrated JAWS frequency response*

Preliminary calibrated measurements of the JAWS amplitude frequency response are presented in Fig. 5. The solid and dotted curves in Fig. 5a correspond to calibrated measurements with and without low-pass filters on the VNA port 2 receivers, respectively. For the case where no filters were used, a ripple of 9 dB was observed in the calibrated data (Fig. 5a). This ripple is caused by calibration artifacts because of a strong distortion in the receiver for the forward wave on port 2. Filtering only this receiver substantially improved the calibrated results, but these results still presented spurs consistent with distortion in the receiver for the backward wave. By low-pass filtering both receivers, we improved the VNA dynamic range and improved the calibrated measurements of the JAWS frequency response. The closeup in Fig. 5b shows a 0.9 dB roll-off and a small ripple. The ripple, which presents a defined oscillating frequency, is believed to be caused by an impedance mismatch between the JJ circuit and the VNA measurement setup.

IV. Conclusions

We have produced the first calibrated measurement of a JAWS system up to 1 GHz. We achieved a calibrated JAWS amplitude frequency response measurement that is within 0.9 dB of the expected amplitude level predicted in [4]. The roll-off is believed to be caused by distortion of the pulse driving signal throughout the radio channel and the remaining ripple is attributed to impedance mismatch between the JJ circuit and our VNA measurement setup.



(a)

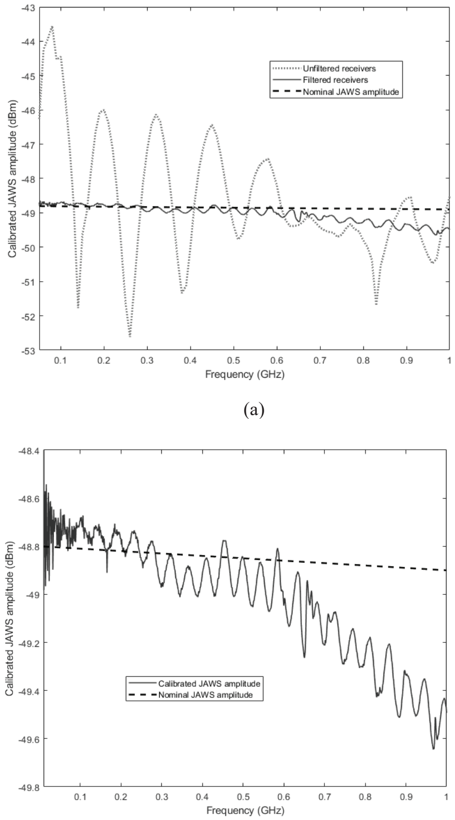
(b)

Fig. 4a) 50 GHz bandwidth measurement of the spectra of a delta-sigma-modulated pattern corresponding to a synthesis frequency of 500 MHz with and without low-pass filters on the VNA port 2 receivers. b) First 10 GHz of the spectrum shown in Fig. 5a. The inset illustrates the intermodulation distortion mechanism in the VNA receivers.

Preliminary load-pull measurements [9] have shown promise for helping to accurately model the JJ circuit (including its frequency-dependent amplitude and impedance). This will allow us to apply mismatch correction to our measurements and resolve the ripple issue. Calibrated broadband measurements of the input bias and quantized output pulses will be of interest to investigate the roll-off in the JAWS frequency response. Future work will also focus on scaling up the frequency and power capabilities of our JAWS system.

V. Acknowledgments

This research was supported by NIST's Innovations in Measurement Science program. We thank Nathan Flowers for his work in quantum voltage standards, Adam Sirois for his work on the cryogenic probe system, Rich Chamberlin for his early work on the cryogenic calibration and the NIST Boulder microfabrication facility and the NIST staff who support it. This work is a contribution of the U.S. Government and is not subject to U.S. copyright.



(a)

(b)

Fig. 5a) Preliminary results of calibrated JAWS amplitude frequency response. b) Closeup of the frequency response.

References

[1] S. P. Benz et al., “A pulse-driven programmable Josephson voltage standard,” Applied Physics Letters, vol. 68, no. 22, pp. 3171–3173, 1996.

[2] O. Kieler et al., "Optical Pulse-Drive for the Pulse-Driven AC Josephson Voltage Standard," in IEEE Transactions on Applied Superconductivity, vol. 29, no. 5, pp. 1-5, Aug. 2019, Art no. 1200205.

[3] J. A. Brevik et al., “Josephson arbitrary waveform synthesis with multilevel pulse biasing,” IEEE Transactions on Applied Superconductivity, vol. 27, no. 3, pp. 1–7, April 2017.

[4] J. A. Brevik et al., “Cryogenic Calibration of a Quantum-based RF Josephson Junction Source”, For publication in ARFTG 2020.

[5] C. A. Donnelly et al., “1 GHz waveform synthesis with Josephson junction arrays,” IEEE Transactions on Applied Superconductivity, vol. 30, no. 3, pp. 1–11, April 2020.

[6] A. S. Boaventura et al., “Traceable characterization of broadband pulse waveforms suitable for cryogenic Josephson voltage applications,” in 2018 IEEE/MTT-S International Microwave Symposium - IMS, June 2018, pp. 1176–1179.

[7] A. S. Boaventura et al., "Microwave Modeling and Characterization of Superconductive Circuits for Quantum Voltage Standard Applications at 4 Kelvin," in IEEE Transactions on Applied Superconductivity. doi: 10.1109/TASC.2019.2963403.

[8] NIST Microwave Uncertainty Framework Calibration Software https://www.nist.gov/services-resources/software/wafer-calibration-software

[9] A. S. Boaventura et. al., “Cryogenic Characterization of Quantum-based RF Waveform Synthesizers for Wireless Communications Metrology”, To be submitted to IEEE Transactions on Applied Superconductivity.

Characterization of a Josephson Junction Comb Generator

Akim A. Babenko[#,$1], Alírio S. Boaventura[#2], Nathan E. Flowers-Jacobs[#3], Justus A. Brevik[#],
Anna E. Fox[#], Dylan F. Williams[#], Zoya Popović[$], Paul D. Dresselhaus[#], Samuel P. Benz[#]
[#]National Institute of Standards and Technology, Boulder, USA
[$]Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, USA
[1]akim.babenko@nist.gov, [2]aliriodejesus.soaresboaventura@nist.gov, [3]nathan.flowers-jacobs@nist.gov

***Abstract*— We present a new type of microwave frequency combs with a potentially calculable pulse shape. The device is an array of 1500 Josephson junctions (JJs) connected in series along a transmission line. The pulse generation is based on the nonlinearity of the JJs. A large-signal network analyzer and a cryogenic probe station are used to characterize the pulses in the frequency domain up to 50 GHz. We compare the measured data to simulations that use the resistively and capacitively shunted JJ model. The amplitude stability of the demonstrated comb generator is better than 0.5 dB per 0.1 dB input drive variation within the operating range. Finally, we observe qualitative agreement between the measured and simulated power spectrum dependence on the input power, and discuss possible improvements to the system model.**

***Keywords*— Pulse measurements, Josephson junctions, Superconducting microwave devices, Millimeter wave measurements, Superconducting integrated circuits.**

I. Introduction

With the development of modern telecommunication systems, there is a high demand for accurate characterization of broadband signals and circuits. Among the most utilized instruments for the analysis of wide-band signals and networks are large-signal network analyzers (LSNAs) [1]. LSNAs provide traceable scattering-parameter, absolute power, and cross-frequency phase measurements.

Currently, a broadband power-meter calibrated using a calorimetric technique provides the traceability for power calibration [2]. For traceable cross-frequency phase calibration, various implementations of comb generators (also referred to as frequency combs) based on semiconductor technology are used, for example split signal pulse generators [3], nonlinear transmission lines [4] and step rocevery diodes [5]. These frequency combs provide stable pulses with rich harmonic content, but the pulse shapes (and therefore frequency content) are not predictable *a priori*. Therefore, comb generators must be calibrated using traceable optical techniques described in detail in [6], where the amplitude and phase stability per 0.1 dB input drive variation was ± 0.6 dB and ± 1.9°, respectively.

In this paper we show the first steps towards creating a comb generator with intrinsically calculable harmonic content that could replace the traceability chain described above. This comb generator is based on Josephson junctions (JJs), which are cryogenically cooled devices made with a thin non-superconducting barrier between two superconductors [7]. When current-biased, a JJ generates pulses with rising and falling edges determined by junction parameters and with a time-integrated voltage across the JJ that is exactly determined by the ratio of defined fundamental constants $\mathrm{h}/2\mathrm{e}$ (Planck constant and the electron charge). This effect is also utilized for the realization of dc [8] and ac [9] voltage standards. Here we study a Josephson junction comb generator similar to the theoretical proposal in [10], which focused on using pairs of JJs connected in parallel to form SQUIDs (Superconducting Quantum Interference Device) biased using magnetic fields.

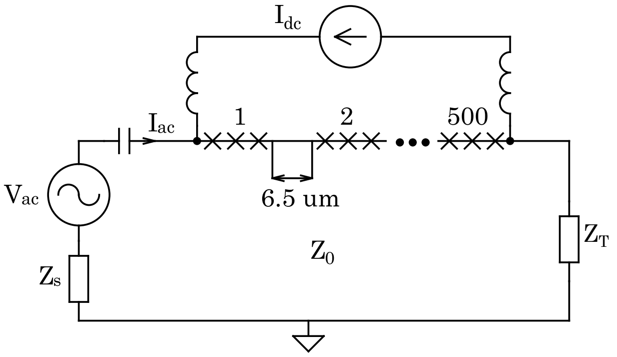


Fig. 1. The diagram of the comb generator. The array of 1500 JJs, placed on a transmission line with the characteristic impedance $\mathrm{Z_0}$, is arranged as 500 stacks with three JJs in each stack. A CW voltage source $\mathrm{V_{ac}}$ drives the junctions with the current $\mathrm{I_{ac}}$, in addition to dc bias $\mathrm{I_{dc}}$.

An array of JJs connected in series, as in Fig. 1, is biased with a combination of dc and 5 GHz CW current. The harmonic content of the resulting voltage pulses across the junctions is measured up to 50 GHz. The calibration and measurements are performed with a cryogenic probe station [11]. Although the exact knowledge of the integrated area is not sufficient to predict the pulse shape, we use a simple model of the JJ, assuming no parameters variation along the array, as well as the ideal behavior of each of the junctions and a lossless substrate, to calculate the expected harmonic content. We show the agreement between these predictions and calibrated measurements, concentrating on the power spectrum stability of the JJ comb generator versus changes in the drive signal amplitude. We also show the initial results for the phase stability of the JJ comb generator.

II. Pulse Generation Technique

The circuit model of a single JJ with the corresponding calculated time-domain response to a 5 GHz sinusoidal input current is shown in Fig. 2.

Babenko, Akim; Soares Boaventura, Alirio De Jesus; Flowers-Jacobs, Nathan; Brevik, Justus; Fox, Anna; Williams, Dylan; Popovic, Zoya;
Dresselhaus, Paul; Benz, Samuel P. "Characterization of a Josephson Junction Comb Generator." Presented at 2020 IEEE MTT-S International
Microwave Symposium (IMS). August 04, 2020 - September 30, 2020.

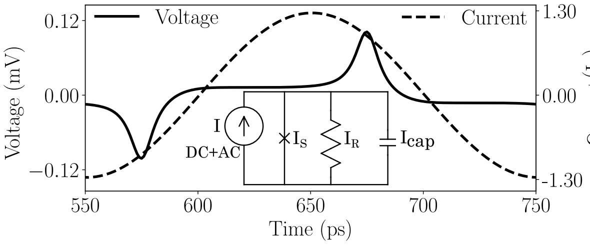


Fig. 2. Simulated time-domain waveforms across a single JJ driven with a 5 GHz CW current. The junction generates pulses with an exact time-integrated voltage area h/2e across the JJ. The pulses are locked to the phase of the drive current.

The dynamics of operation are modeled using the widely-accepted resistively and capacitively shunted junction (RCSJ) equivalent electrical circuit [7], as shown in Fig. 2. The drive current I divides into the superconducting (I_S), resistive (I_R) and capacitive (I_{cap}) currents where $I_{cap} \ll I_R, I_S$ for the technology described below. If the applied current I is below the critical current value, I_C, it flows solely into the superconducting branch and obeys the dc Josephson equation $I = I_S = I_C \sin\phi$, where ϕ is the quantum-mechanical macroscopic phase difference across the JJ [7]. This zero-voltage state corresponds roughly to the flat regions of voltage waveform in Fig. 2. When the input current I exceeds the critical value I_C, a non-zero voltage V evolves across the junction, as given by:

$$I = I_C \sin\phi + V/R_n + C \cdot \partial V/\partial t, \tag{1}$$

$$\partial\phi/\partial t = (2e/\hbar)V, \tag{2}$$

where (2) is the ac Josephson equation. Substituting (2) into (1), one obtains a second-order nonlinear differential equation for the phase difference ϕ and solves for V as a function of ϕ. The solution under dc, CW, or pulse drive current $I > I_C$ is of the form of voltage pulses with integrated area equal exactly to $\pm h/2e$. In the dc-only bias case, however, the timing of these pulses is difficult to control and the circuit is sensitive to current noise. Therefore, the voltage pulses are typically generated with a stabilized CW or pulse drive current. An example is shown in Fig. 2, with exactly one pulse per half-period of the CW input drive. The range of bias parameters over which an integer number of output voltage pulses is locked to the input drive is referred to as the Quantum Locking Range (QLR) [9]. The phase shift of the voltage with respect to the drive current stems from (1), and for small input signals $I \ll I_C$ the JJ can be approximated as a non-linear inductor [7]. It is important to note again that the voltage pulses are always phase-locked to the drive current.

When the JJ capacitance C is negligibly small, as in this paper, the –3 dB cutoff frequency is determined as $f_c = I_C R_n (2e/h)$, setting the rising and falling edges of the voltage pulses. Thus, the harmonic content of those voltage pulses is set mainly by the intrinsic properties of the junctions and not by the frequency content of the drive signal, particularly for signals with bandwidth $\ll f_c$. The 0.12 mV amplitude, shown in Fig. 2, was calculated for the JJs in this paper with measured $I_C R_n = 53\,\mu$V and $f_c = 26$ GHz. From the right y-axis of the plot, one can also see that the amplitude of the 5 GHz drive current should be about 30% above the I_C value to maintain the QLR with exactly one voltage pulse per half-period. The small pulse amplitude across a single JJ is the motivation for connecting thousands of junctions in series to increase the total output voltage.

III. Simulation Setup

We simulated our comb generator using the circuit in Fig. 3 with the WRSpice [12] open-source transient circuit simulator, which includes the RCSJ model of a JJ. The large inductors at the ends of the JJ array are used to both supply a dc current bias to the JJs and to measure the dc and low-frequency content of the voltage pulses. The LSNA CW signal source is modeled as a 5 GHz voltage source with an impedance Z_S that is measured during the calibration procedure described below. The large capacitor is part of the LSNA's dc bias tees. As in Fig. 1, we arrange the series-connected JJs in groups of three that are separated by 6.5 μm along a 50 Ω lossless transmission line. We consider the silicon substrate as a lossless dielectric as our first approximation. The impedance of the termination resistor is determined using a calibrated measurement of a separate on-chip termination structure and is equivalent to a series RL-network with $R_T = 59\,\Omega$ and $L_T = 96$ pH.

The WRSpice solver solutions are the total voltage V_1 and current I_1 as functions of time at the input node of and through the array, respectively. Those are converted into the peak complex amplitudes of the a_1 and b_1 power waves [13], as shown in Fig. 3 and given by:

$$a_1 = (V_1 + I_1 Z_0)/(2\sqrt{|Z_0|}), \tag{3}$$

$$b_1 = (V_1 - I_1 Z_0)/(2\sqrt{|Z_0|}), \tag{4}$$

where Z_0 is equal to 50 Ω. The a_1 wave contains two superimposed components: the incident drive signal component a_1^{inc} and multiple-reflected component a_1^{ref} stemming from the mismatch between the source and the array input impedances. It is conceptually useful to consider the b_1 wave in two different current bias regimes. When the current bias is $I < I_C$, then b_1 contains two distinct components: the linear reflected component due to the impedance mismatch b_1^{mm} and the component b_1^{nl} generated by the small-signal non-linear inductance of the JJs which creates harmonics but not pulses. However, when the current bias is $I > I_C$ then b_1 contains the same impedance mismatch term b_1^{mm} but the

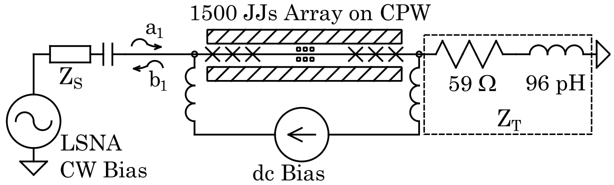


Fig. 3. The circuit used to model the response of the JJ comb generator to a CW drive signal. The values of the source Z_S and termination Z_T impedances are obtained from prior scattering-parameter measurements on the same chip.

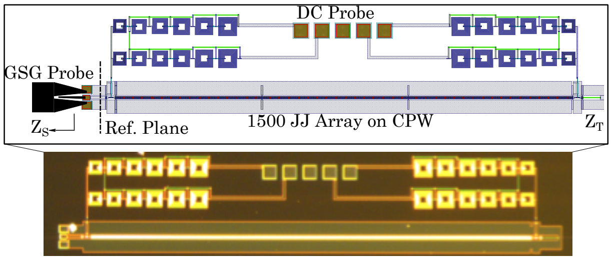


Fig. 4. The layout (top) and photo (bottom) of the comb generator chip. The total length of the array $l = 3.25$ mm on the 350 μm thick silicon substrate ($\epsilon_r = 12$) satisfies the lumped-element limit $l < \lambda/8$ for frequencies up to 5 GHz. The CPW consists of a 16 μm wide center conductor with an 8 μm gap to fulfill the 50 Ω requirement.

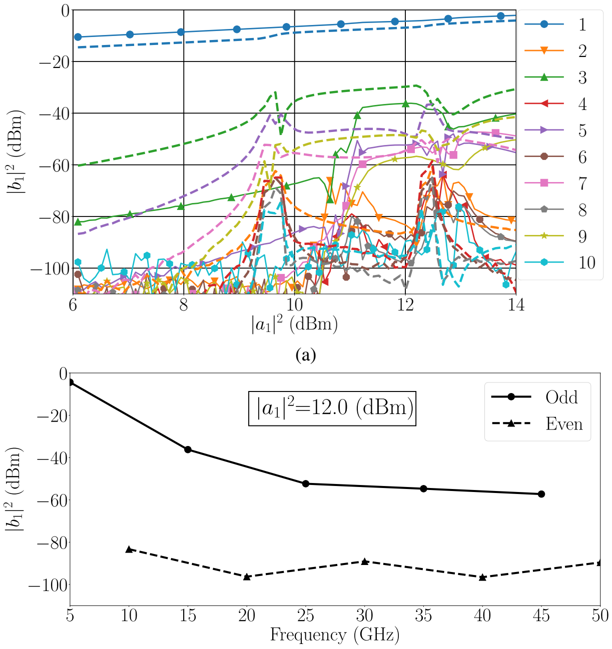
non-linear term is now better understood as JJ pulses b_1^{pp} propagating towards the generator.

IV. Chip Layout and Measurement Procedure

The layout (top) and photo (bottom) of the JJ comb generator chip are shown in Fig. 4. The chip is fabricated with niobium superconducting electrodes and $\mathrm{Nb}_x\mathrm{Si}_{1-x}$ normal-metal barriers described in detail in [14]. The 1500 junctions are embedded in a superconducting copalanar waveguide (CPW) line with Nb signal and ground conductors on an oxidized silicon substrate. The characteristic impedance is designed to be 50 Ω and the line is terminated with an on-chip resistor. The array of series-connected junctions consists of 500 vertical stacks of 3 JJs distributed along the central conductor of the CPW with one stack every 6.5 μm, which implies that each stack can be treated as a lumped-element circuit for the relevant measurement frequencies below 50 GHz.

The circuit is probed with a 150 μm pitch ground-signal-ground (GSG) probe with a cryogenic probe station. The temperature of the chip is set to 4 K. The critical current is temperature dependent; at 4 K we determined that $I_C = 13.9\,\mathrm{mA}$ and $R_n = 3.8\,\mathrm{m}\Omega$ based on measurements of the dc current-voltage characteristics [7]. As in the simulation setup, the spiral-inductor coils are used for dc and low-frequency biasing and measurements along with a 5-pin dc probe.

A two-tier wave-parameter calibration of the LSNA was performed to move the reference plane of the measurements to the input of the JJ array [11], as in Fig. 4. For the first tier, we performed a two-port Short-Open-Load-Thru (SOLT) calibration at the plane of the room-temperature inputs to the cryogenic probe station. At the same plane, we also performed an absolute power and cross-frequency phase calibrations using a power meter and a commercial comb generator, respectively. The second tier calibration used on-chip superconducting calibration structures to perform a Multiline Thru-Reflect-Line calibration, renormalizing the obtained reference impedance to 50 Ω. The frequency grid for the calibration was from 1.25 GHz to 50 GHz with a 1.25 GHz step. We also set the LSNA noise floor to –110 dBm by choosing a 10 Hz intermediate frequency bandwidth.



(a)

(b)

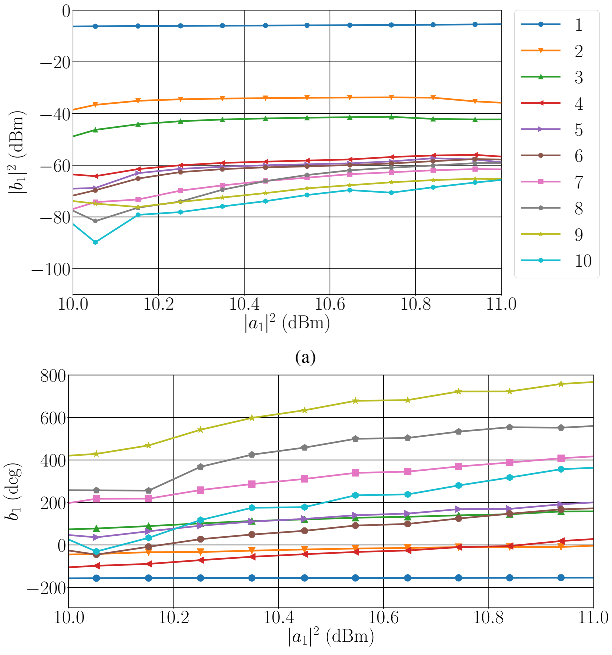
Fig. 5. Pulse power: (a) The measured (solid, with markers) and simulated (dashed, no markers) power of the harmonics in the pulses traveling towards the source versus the power of the incident 5 GHz drive signal. The even simulated harmonics are also small and therefore not shown. (b) The measured power of the harmonics versus frequency for a 12 dBm input power.

V. Measured and Simulated Results

To demonstrate the performance of our comb generator, we first test the change in output power as a function of input drive amplitude. We follow [3] and first focus on the generation of symmetric pulses. As in Fig. 2, a 5 GHz CW input drive signal from the LSNA and no dc bias are applied to the array. The input drive signal power is swept from 6 dBm to 14 dBm at the reference plane.

The measured and simulated results are shown in Fig. 5, where it is clear that the fundamental component of the b_1-wave has the mismatch term, b_1^{mm}, which dominates. Based on the lowest input power levels, we determine that the return loss of the circuit at the fundamental frequency is 16.6 dB. For the second and third harmonics, non-zero b_1^{nl} components exist for the input power less than about 10.5 dBm due to the small-signal non-linearity of the junctions. For the third and higher-order odd harmonics, note a QLR between 11.2 dBm and 12.1 dBm where the pulse component b_1^{pp} dominates, as in Fig. 2. As expected for symmetrical pulses in this range, the even harmonics are suppressed by greater than 30 dB relative to the neighboring odd harmonics. Within the QLR, the third harmonic at 15 GHz varies by 0.2 dB per 0.1 dB change in the input power. The variation increases with frequency, and the ninth harmonic at 45 GHz varies by 0.5 dB per 0.1 dB change in the input power. Above an input power of about 13 dBm we also begin to observe the generation of two pulses per half-period of the drive signal.

Babenko, Akim; Soares Boaventura, Alirio De Jesus; Flowers-Jacobs, Nathan; Brevik, Justus; Fox, Anna; Williams, Dylan; Popovic, Zoya; Dresselhaus, Paul; Benz, Samuel P. "Characterization of a Josephson Junction Comb Generator." Presented at 2020 IEEE MTT-S International Microwave Symposium (IMS). August 04, 2020 - September 30, 2020.



(a)

(b)

Fig. 6. The measured power (a) and phase (b) of the harmonics in the pulses traveling towards the source with a positive dc bias and a 5 GHz CW input drive. The phases of the harmonics are normalized to that of the fundamental component of the incident wave a_1 [1].

We have identified two main simplifications in the model which lead to the differences between the simulations and measurements. First, the differences in the nonlinear small-signal response likely occur because the JJs do not follow a simple sinusoidal dc Josephson equation and instead have a more complicated dependence on ϕ [15]. Second, fabrication tolerances cause variations in the I_c and R_n values of the JJs within the array which is not yet included in our model. The JJ non-uniformity is likely why the simulated QLR is wider than the measured QLR and why the simulated QLR starts at 1 dB smaller input drive power.

Finally, we show the measured power and phase spectra versus input signal amplitude in Fig. 6, where we apply a positive 4 mA dc bias so that only positive pulses are generated. The same level of the power spectrum variation is seen as for the case of symmetric pulses. There is a linear trend in the phase variation and the slope increases with increasing harmonic number n. Specifically, we observe an approximate slope of $8.5 + 4.5(n - 2)$ degrees per 0.1 dB increase in input amplitude. This trend was consistently obtained in several consecutive measurements.

VI. Conclusion

In this paper, we show the characterization of a JJ comb generator that generates stable pulses with a predictable harmonic content. We also compare the power of the measured harmonics to simulations involving a simple JJ model, explain the input drive amplitude dependence observed in the measurements, and discuss how the model can be improved. Finally, we achieve an amplitude stability versus input drive variation comparable to that in [6]. Future work will be devoted to more accurately simulating the JJ arrays, accounting for the mismatch term in the fundamental component, investigating in more detail the phase spectrum, characterizing a two-port JJ comb generator, increasing the pulse amplitude by increasing the number of JJs in the array, and improving the comb frequency response by increasing the cutoff frequency of the individual JJ. We will also address the cases of lower-frequency drive signals and thus lower pulse repetition frequencies in future measurements.

Acknowledgment

The authors would like to thank Adam Sirois and Manuel Castellanos-Beltran from NIST for their support with the experimental setup and the NIST Boulder Microfabrication Facility for assistance in fabricating the superconducting chips used in this work. This work is a contribution of the U.S. government and is not subject to U.S. copyright.

References

[1] J. Verspecht, “Large-signal network analysis,” *IEEE Microwave Magazine*, vol. 6, no. 4, pp. 82–92, Dec 2005.

[2] R. A. Ginley, “Traceability for microwave power measurements: Past, present, and future,” in *2015 IEEE 16th Annual Wireless and Microwave Technology Conference (WAMICON)*, April 2015, pp. 1–5.

[3] D. B. Gunyan and J. B. Scott, “Pulse generator,” U.S. Patent 7 423 470B2, Sep. 9, 2008.

[4] M. J. W. Rodwell, D. M. Bloom, and B. A. Auld, “Nonlinear transmission line for picosecond pulse compression and broadband phase modulation,” *Electronics Letters*, vol. 23, no. 3, pp. 109–110, January 1987.

[5] H. T. Friis, “Analysis of harmonic generator circuits for step recovery diodes,” *Proceedings of the IEEE*, vol. 55, no. 7, pp. 1192–1194, July 1967.

[6] H. C. Reader, D. F. Williams, P. D. Hale, and T. S. Clement, “Comb-generator characterization,” *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 2, pp. 515–521, Feb 2008.

[7] T. V. Duzer and C. W. Turner, *Principles of Superconductive Devices and Circuits*, 2nd ed. Upper Saddle River, N.J: Prentice Hall PTR, 1999, ch. 4.

[8] C. J. Burroughs *et al.*, “NIST 10 V programmable josephson voltage standard system,” *IEEE Trans. Instrum. Meas.*, vol. 60, no. 7, pp. 2482–2488, July 2011.

[9] N. E. Flowers-Jacobs *et al.*, “Two-volt Josephson arbitrary waveform synthesizer using Wilkinson dividers,” *IEEE Trans. Appl. Supercond.*, vol. 26, no. 6, pp. 1–7, Sep. 2016.

[10] P. Solinas, S. Gasparinetti, D. Golubev, and F. Giazotto, “A Josephson radiation comb generator,” *Nature Scientific Reports*, vol. 5, no. 12260, pp. 1244–1245, 2015.

[11] A. S. Boaventura *et al.*, “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, Mar. 2020.

[12] WRspice Circuit Simulations. Whiteley Res. Sunnyvale, CA, US. [Online]. Available: http://wrcad.com

[13] D. Williams, “Traveling waves and power waves: Building a solid foundation for microwave circuit theory,” *IEEE Microw. Mag.*, vol. 14, no. 7, pp. 38–45, Nov 2013.

[14] B. Baek, P. D. Dresselhaus, and S. P. Benz, “Co-sputtered amorphous Nb$_x$ Si$_{1-x}$ barriers for josephson-junction circuits,” *IEEE Trans. Appl. Supercond.*, vol. 16, no. 4, pp. 1966–1970, Dec 2006.

[15] A. A. Golubov, M. Y. Kupriyanov, and E. Il’ichev, “The current-phase relation in josephson junctions,” *Rev. Mod. Phys.*, vol. 76, pp. 411–469, Apr 2004. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.76.411

Preprint: 9th International Workshop on Combinatorial Testing (IWCT 20), Porto, Portugal, October 24-28, 2020.

Combinatorial Methods for Explainable AI

D. Richard Kuhn[1], Raghu N. Kacker[1], Yu Lei[2], Dimitris E. Simos[3]

*[1] National Institute of
Standards and Technology
Gaithersburg, MD 20899, USA
{kuhn, raghu.kacker}@nist.gov*

*[2]Computer Science & Engineering
University of Texas at Arlington
Arlington, TX, USA
ylei@uta.edu*

*[3]SBA Research
Vienna, Austria
dsimos@sba-research.org*

Abstract **–This short paper introduces an approach to producing explanations or justifications of decisions made by artificial intelligence and machine learning (AI/ML) systems, using methods derived from fault location in combinatorial testing. We use a conceptually simple scheme to make it easy to justify classification decisions: identifying combinations of features that are present in members of the identified class and absent or rare in non-members. The method has been implemented in a prototype tool, and examples of its application are given.**

Keywords – artificial intelligence; combinatorial testing; explainable AI; machine learning; t-way testing;

I. Introduction

Artificial intelligence and machine learning (AI/ML) systems have exceeded human performance in nearly every application where they have been tried and are increasingly incorporated into consumer products. As the current trend continues, AI will be increasingly used in safety-critical systems such as self-driving cars, medical devices, and weapons systems. Current AI systems are generally accurate, but sometimes make mistakes, and human users will not trust their decisions without explanation. Consequently, there is a significant need for improvements in explainability of AI/ML system functions and decisions [1][2][3][4][5][6][7].

The central problem for explainability, according to the Defense Advanced Research Progress Agency (DARPA), is to provide sufficient justification for an AI/ML conclusion such that users know why a conclusion was reached, or why not, and to allow the user to know when an algorithm will succeed or fail, and when it can be trusted [1]. Many conventional approaches leave users wondering what inputs caused a particular conclusion. More than curiosity is involved, as many AI/ML applications may be safety critical, and accuracy rates that are high enough for some applications are inadequate when safety and lives are at risk. Analysis within the aerospace industry concludes that the “artificial intelligence (AI) technology that has made spectacular progress in the consumer world is thus far unsuited to air transport safety standards”, and explainability will be essential for certification by regulatory authorities [8]. Ideally, the ML algorithm should be able to explain its conclusion in a manner similar to a human expert, so that other human experts can have confidence in a conclusion, or spot a flaw in the reasoning. This is a significant challenge for methods such as neural networks.

Typically, there is a tradeoff between AI/ML accuracy and explainability: the most accurate methods, such as convolutional neural nets (CNNs), provide no explanations, while more understandable methods, such as rule-based systems, tend to be less accurate [1][2]. Black-box statistical predictions are inadequate, and explanations must be understandable to non-specialists, such as physicians, financial analysts, and in many cases everyday users.

The need for explainability in AI was recognized early, and was an inherent component of many of the first AI diagnostic systems. These were often expert systems using programming-style if/then rules to make decisions. For example: “if patient has symptoms A and B, or has B with C and D, then illness is X”. Such systems provide natural explanations, but rules can be difficult to identify, and in many cases are less accurate than other approaches.

While neural networks and related methods often provide better accuracy, they are opaque to users. Decisions are produced using a vast number of internal connections, and some efforts have been aimed at adding explanations to neural nets, but this is an ongoing area of research and the approach has not been widely adopted.

A third way of adding explanations is model induction, inferring an explainable model from black-box inputs and outputs. Systems using this approach have been produced to attempt to identify the most relevant features used in decisions, typically using statistical methods. For example, LIME, one of the more widely used methods, determines features that are most strongly associated with an output [7]. Our method is most similar to these systems, except that we identify combinations of feature values. The result is a system that can infer explanations that incorporate predicates similar to rule based systems, using input/output combinations. Predicates that identify distinguishing features can also assist in validating the model generated by the ML algorithm, by providing more information for human experts in validation. Thus, this method adds value in AI/ML for both users, who need explanation, and model developers seeking to validate a black box model.

II. Human Factors Aspects

A key question for explainability is the degree to which an explanation will be acceptable and trusted by users, which necessarily deals with both technical and human factors. Full development of explainable AI will require extensive validation through human testing, which has not been included in most work in the field [9]. However, the applicability of human factors research to explainable AI has been studied, building on extensive research from psychology on models of human explanation. Here we summarize the major findings of this work, as documented in surveys of the field [9] [10][11][12] [13][14].

Miller et al. [10] suggest that “the most important result from this work is that explanations are *contrastive*: or more

Kuhn, D. Richard; Kacker, Raghu N.; Lei, Yu; Simos, Dimitris. ”Combinatorial Methods for Explainable AI.” Presented at 2020 IEEE 13th International Conference on Software Testing, Verification and Validation Workshops (ICSTW 2020), Porto, PT. October 24, 2020 - October 28, 2020.

accurately, *why-questions* are contrastive. That is, why-questions are of the form 'Why P rather than Q?'" They note that the psychological research indicates that generally all why-questions are contrastive, seeking an implicit contrast case even if one is not stated explicitly. This is suggested as a potential approach to explainable AI, because providing a contrast case may be easier than a full set of causes [10][15].

Another aspect of explainability identified in human factors research is causal attribution, i.e., the manner in which causes are attributed to events. Relevant findings in this area include research showing that users may consider counterfactuals, what would have happened if some event was not present [16][17], and that only a subset of a full event chain is typically used [17]. It is perhaps not surprising that users prefer simple explanations, with fewer causes or factors, but it was also found that simpler explanations were preferred over more likely explanations [18].

III. Method

The classification problem in machine learning is in some ways similar to the problem of fault location in combinatorial testing for software. The objective in both cases is to identify a small number of interactions, out of possibly billions or more, that trigger a failure (in testing) or produce a conclusion (in machine learning). We have methods and tools for fault location in combinatorial testing that can be adapted to ML problems, to identify the rare combinations of variable values that produce conclusions in AI systems. This approach has not been applied to explainable AI before.

A basic approach to fault location for testing is to subtract the set of combinations in passing tests from the set of combinations in failing tests, then using appropriate strategies to narrow down the remaining set to the most likely failure inducing combinations. Similar strategies involve identifying combinations that are more common in failing than in passing tests, to find the most likely cause of a failure.

We can apply this general strategy to the AI/ML explanation problem. Suppose a given object has been identified as a member of a particular class, one of the most common operations in machine learning. A vast number of algorithms and statistical methods can be used to make this identification, but it must be explained to users why the object belongs to the selected class and not some other class. That is, what inputs to the classification algorithm are a convincing justification for concluding that the object is in class X and not any other class? This is very similar to the problem of determining what inputs are the reason for a test to produce a failure rather than a passing result.

For explainability, we will also consider two sets of combinations – class and non-class member features, where 'class' refers to a particular group that an object is assigned to. For example, as illustrated in Fig. 1, we may want to explain why an animal is classified as a cat, noting that it shares features with other class members - brown & furry, whiskers, claws – and it does not have features of animals outside the cat class - not aquatic, not venomous. Some features are shared by both the class and non-class members.

While a variety of statistical methods are available for identifying one or a few features that contribute to a conclusion, more information can be provided by using methods from combinatorial testing fault location. We will consider *t*-way combinations, seeking to identify combinations that are unique to class members, i.e., not present in non-class members. It is likely that single features will not be unique, and many 2-way or higher strength feature combinations will also not be unique. But by considering *t*-way combinations with increasing values of *t*, we are likely to reach a point where some t-way combinations are uniquely associated with the class under consideration, or are never associated with the class and can be used to exclude it.

Individual features (orange)
– brown & furry, whiskers, claws, not aquatic, not venomous, 4 legs, **...**

Class features (yellow) -
brown & furry, black & furry, whiskers, claws, **...**not aquatic, not venomous, 4 legs,

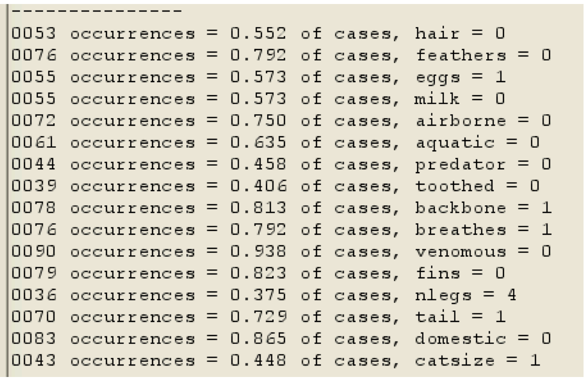
Fig. 1. Feature identification

Looking at combinations of features makes sense intuitively for explanations, because individual features are normally too widely shared among objects of different types. Among animals, thousands of types have four legs, or claws, or pointed ears, but only a limited number have all of these features. For explanations we will look for combinations that are unique, or extremely rare. This is of course essentially the same process used to identify members of a taxonomy, by looking for features an object shares with members of a defined class and for other features that exclude it from a specific class. Thus we argue that the method introduced here is intuitive for users. By adapting combinatorial fault location processes, we can enhance and improve this intuitive method, by considering huge numbers of feature combinations, and quantifying their degree of association with members/non-members of classes. In the following section we illustrate the effectiveness of this approach with an example.

Example. For a more comprehensive example, and to illustrate the application of a prototype tool referred to as ComXAI, we will use the Animals with Attributes (AwA) database to explain the classification of an animal as a reptile. The AwA database describes a large collection of animals using 16 features, 15 boolean and one with six values. For example, Testudo the tortoise [22] (University of Maryland mascot), is shown in Fig. 2, with the following attributes (where 0=false, 1=true): *hair*=0, *feathers*=0, *egg-laying*=1, *milk-producing*=0, *airborne*=0, *aquatic*=0, *predator*=0, *toothed*=0, *backbone*=1, *breathes*=1 *venomous*=0, *fins*=0, *num-legs*=4, *tail*=1, *domestic*=0, *cat-size*=1.

Suppose that an AI/ML algorithm has assigned the class reptile to Testudo. The prototype ComXAI tool analyzes the presence of combinations of these features in the AwA database animals that are not reptiles. The objective is to identify combinations of reptile features, present in Testudo, that are not present (or extremely rare) in non-reptiles. The presence of these feature combinations should be sufficiently convincing that a reptile has been identified correctly.

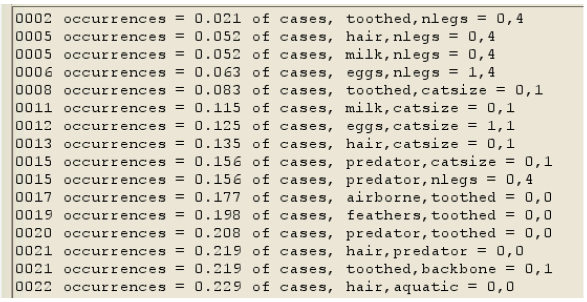
Fig 2. Why is this creature recognized as a reptile?



```
---------------
0053 occurrences = 0.552 of cases, hair = 0
0076 occurrences = 0.792 of cases, feathers = 0
0055 occurrences = 0.573 of cases, eggs = 1
0055 occurrences = 0.573 of cases, milk = 0
0072 occurrences = 0.750 of cases, airborne = 0
0061 occurrences = 0.635 of cases, aquatic = 0
0044 occurrences = 0.458 of cases, predator = 0
0039 occurrences = 0.406 of cases, toothed = 0
0078 occurrences = 0.813 of cases, backbone = 1
0076 occurrences = 0.792 of cases, breathes = 1
0090 occurrences = 0.938 of cases, venomous = 0
0079 occurrences = 0.823 of cases, fins = 0
0036 occurrences = 0.375 of cases, nlegs = 4
0070 occurrences = 0.729 of cases, tail = 1
0083 occurrences = 0.865 of cases, domestic = 0
0043 occurrences = 0.448 of cases, catsize = 1
```

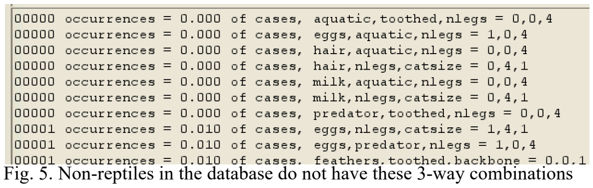
Fig. 3. non-reptile single feature combinations.

As shown in Fig. 3, no single feature is sufficient explanation for classifying Testudo as a reptile, as he shares features with non-reptiles. For example, 55.2 % of other animals in the database have no hair, and 79.2 % have no feathers. Additionally, no pair of features is sufficient, as Testudo also shares 2-way combinations with non-reptiles. As shown in Fig. 4, 2.1 % of the animals in the AwA database have the feature pair *toothless & four-legged*, and 5.2 % have the feature pair *milk-producing & four-legged*.



```
0002 occurrences = 0.021 of cases, toothed,nlegs = 0,4
0005 occurrences = 0.052 of cases, hair,nlegs = 0,4
0005 occurrences = 0.052 of cases, milk,nlegs = 0,4
0006 occurrences = 0.063 of cases, eggs,nlegs = 1,4
0008 occurrences = 0.083 of cases, toothed,catsize = 0,1
0011 occurrences = 0.115 of cases, milk,catsize = 0,1
0012 occurrences = 0.125 of cases, eggs,catsize = 1,1
0013 occurrences = 0.135 of cases, hair,catsize = 0,1
0015 occurrences = 0.156 of cases, predator,catsize = 0,1
0015 occurrences = 0.156 of cases, predator,nlegs = 0,4
0017 occurrences = 0.177 of cases, airborne,toothed = 0,0
0019 occurrences = 0.198 of cases, feathers,toothed = 0,0
0020 occurrences = 0.208 of cases, predator,toothed = 0,0
0021 occurrences = 0.219 of cases, hair,predator = 0,0
0021 occurrences = 0.219 of cases, toothed,backbone = 0,1
0022 occurrences = 0.229 of cases, hair,aquatic = 0,0
```

Fig. 4. 2-way non-reptile feature combinations.



```
00000 occurrences = 0.000 of cases, aquatic,toothed,nlegs = 0,0,4
00000 occurrences = 0.000 of cases, eggs,aquatic,nlegs = 1,0,4
00000 occurrences = 0.000 of cases, hair,aquatic,nlegs = 0,0,4
00000 occurrences = 0.000 of cases, hair,nlegs,catsize = 0,4,1
00000 occurrences = 0.000 of cases, milk,aquatic,nlegs = 0,0,4
00000 occurrences = 0.000 of cases, milk,nlegs,catsize = 0,4,1
00000 occurrences = 0.000 of cases, predator,toothed,nlegs = 0,0,4
00001 occurrences = 0.010 of cases, eggs,nlegs,catsize = 1,4,1
00001 occurrences = 0.010 of cases, eggs,predator,nlegs = 1,0,4
00001 occurrences = 0.010 of cases, feathers,toothed,backbone = 0,0,1
```

Fig. 5. Non-reptiles in the database do not have these 3-way combinations

Looking at 3-way combinations produces much more useful results. As seen in Fig. 5, several 3-way feature combinations uniquely identify reptiles among the animals in the AwA database. No other genus is *non-aquatic & toothless & four-legged*; no other is *egg-laying & non-aquatic & four-legged*, and so on. Only reptiles, among the animals in the database, have the 3-way combinations of features shown in Fig. 5.

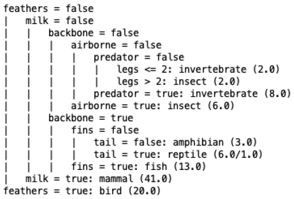
It is important to note that a different picture emerges from simply listing the individual features that are the strongest differentiators: *four legs*, *toothless*, *cat-size*. As seen in Fig. 3, none of these individual features is anywhere near adequate for identifying a reptile. Additionally, the 3-way combination of these individually-identified features does not appear in the list of 3-way combinations that uniquely identify a reptile among animals in the database. There are in fact many animals in the database with the features *four legs & toothless & cat-size.* This is a significant difference between the ComXAI approach and methods of statistically identifying the most significant features individually. This example shows why it is necessary to check the rate of occurrence of *t*-way combinations, rather than assume that the *t* strongest associations individually are sufficient to explain a classification.

IV. Discussion

Validation of this method using human subjects is outside the scope of this work, but we can consider the ComXAI approach with respect to human factors research on explanation, introduced in Sect. II. The method and tool described in this paper have been designed to provide intuitive explanations by identifying *t*-way combinations that are present in a given member of a class, and not present or extremely rare in non-members. We believe this is a natural form of explanation because it relies on observable features but quantifies the degree to which feature combinations occur in the class and non-class sets. In particular, this approach provides explanations that are *contrastive*, often considered the most important characteristic of explanations in the psychological literature [10][11]. We identify combinations of attributes that characterize the class to be identified, and that are not found in non-members of this class. This process naturally produces explanations that are contrastive – the combinations presented in the explanation are uniquely associated with the class identified. This provides a clear answer to the "Why *P* and not *Q*?" question implicit in explanations. The class is *P* because these combinations occur only in *P*, and do not occur with any other class *Q*. Using methods developed for fault location makes it possible to apply the approach across many *t*-way combinations, providing strong justifications for AI/ML conclusions.

It should also be noted that identifying t-way combinations of features that distinguish a class member is essentially the same as specifying predicates in a rule-based expert system. Referring back to Example 1, the six 3-way combinations could be mapped directly to a rule such as "if (*not aquatic && not toothed && four legs*) || (*egg-laying* && *not aquatic* && *four legs*) …. then genus = testudo". It is often suggested that rule-based expert systems are the most interpretable, so this correspondence between *t*-way combinations and rule-based predicates also suggests that the ComXAI explanations can be understood well by users.

This method can also be compared with a decision tree approach, where leaf nodes are *t*-way combinations of features (Fig. 6). Note that the tree uses more attributes, leading to more complex predicates, while ComXAI identifies unique combinations of only three features. We plan to investigate the potential for such decision minimization in the future.



```
feathers = false
|   milk = false
|   |   backbone = false
|   |   |   airborne = false
|   |   |   |   predator = false
|   |   |   |   |   legs <= 2: invertebrate (2.0)
|   |   |   |   |   legs > 2: insect (2.0)
|   |   |   |   predator = true: invertebrate (8.0)
|   |   |   airborne = true: insect (6.0)
|   |   backbone = true
|   |   |   fins = false
|   |   |   |   tail = false: amphibian (3.0)
|   |   |   |   tail = true: reptile (6.0/1.0)
|   |   |   fins = true: fish (13.0)
|   milk = true: mammal (41.0)
feathers = true: bird (20.0)
```

Fig. 6. J48 decision tree for AwA produced by Weka [20].

It is also possible to use combinatorial methods to check for gaps in ML models [21]. This approach might be used in concert with ComXAI to validate ML models.

V. Conclusions

Explainability is a critical problem in the acceptance of artificial intelligence/machine learning, especially for critical applications. Human users may not trust AI if conclusions cannot be explained. Methods from combinatorial testing can be applied to the problem of explainable AI, by determining combinations of variable values that differentiate an example from other possible conclusions. That is, we identify *t*-way combinations that are present in members of a class and not present in objects outside the class. A prototype tool ComXAI that applies this approach has been developed.

Acknowledgement and disclaimer: This paper is an extended version of a NIST technical report [19]. Products may be identified in this document, but such identification does not imply recommendation by NIST, nor that those identified are necessarily the best available for the purpose.

References

[1] Gunning D. Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). 2017. http://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

[2] Biran O, Cotton C. Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI)* 2017 (p. 8).

[3] Brinton C. A Framework for explanation of machine learning decisions. *IJCAI-17 Workshop on Explainable AI (XAI)* 2017 .

[4] Lomas M, Chevalier R, Cross II EV, Garrett RC, Hoare J, Kopack M. Explaining robot actions. Proceedings of the seventh annual *ACM/IEEE international conference on Human-Robot Interaction* 2012 Mar 5 (pp. 187-188). ACM.

[5] Belle V. Logic meets probability: towards explainable AI systems for uncertain worlds. *26th Intl Joint Conference on Artificial Intelligence, IJCAI* 2017

[6] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?. arXiv:1712.09923. 2017 Dec 28.

[7] Shakerin, F., & Gupta, G. (2018). Induction of Non-Monotonic Logic Programs to Explain Boosted Tree Models Using LIME. *arXiv:1808.00629*.

[8] T. Dubois, "No AI in Cockpit Anytime Soon, Onera, Thales Say", *Aviation Week and Space Technology*, Nov. 26, 2018.

[9] Tjoa, E., & Guan, C. (2019). A survey on explainable artificial intelligence (XAI) *arXiv:1907.07374*.

[10] Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioral sciences. *arXiv:1712.00547*.

[11] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intell.*, *267*, 1-38.

[12] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

[13] Zhang, Y., & Chen, X. (2018). Explainable recommendation: A survey and new perspectives. *arXiv:1804.11192*.

[14] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," *2018 41st Intl Convention on Information and Communication Tech., Electronics and Microelectronics (MIPRO)*, Opatija, 2018, pp. 0210-0215.

[15] Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, *27*, 247-266.

[16] Kahneman, D., & Tversky, A. (1981). *The simulation heuristic* (No. TR-5). Standford Univ. DOI or website needed

[17] Hilton, D. J., & JOHN, L. M. (2007). The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking* (pp. 56-72). Routledge.

[18] Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232-257.

[19] Kuhn, R., & Kacker, R. (2019). *An Application of Combinatorial Methods for Explainability in Artificial Intelligence and Machine Learning*. NIST, 5/22/19.

[20] Weka data mining. https://www.cs.waikato.ac.nz/ml/weka/

[21] Barash, G., Farchi, E., Jayaraman, I., Raz, O., Tzoref-Brill, R., & Zalmanovici, M. Bridging the gap between ML solutions and their business requirements using feature interactions. *2019 27th ACM Joint Meeting European Software Eng. Conf* (pp. 1048-1058

[22] https://en.wikipedia.org/wiki/Diamondback_terrapin#/media/File:Testudo_2.jpg

A CASE STUDY FOR MODELING MACHINE TOOL SYSTEMS USING STANDARD REPRESENTATIONS

Maja Bärring[1]; Guodong Shao[2]; Moneer Helu[2]; Björn Johansson[1]

[1]Chalmers University of Technology, Industrial and Materials Science, Hörsalsvägen 7A, Gothenburg, 41296, Sweden

[2]National Institute of Standards and Technology (NIST), 100 Bureau Dr., Gaithersburg, MD 20899, USA

ABSTRACT

Machine models play an important role to support decision making for purchasing, scheduling, and routing in manufacturing. However, it is challenging to share a machine model that is developed using proprietary formats. A model of a fully assembled machining system in a neutral format can help overcome this challenge. Standard-based machine tool models will not only facilitate information reuse but also enable model exchange between systems. In this paper a case study is discussed to demonstrate the initial effort of a standard representation for a machining system including both component geometric and kinematics information. This standard-based machine model will be easily imported to another tool.

Keywords –CAx tools, kinematics, interoperability, machine model, standards, smart manufacturing systems, STEP

1. INTRODUCTION

Smart manufacturing systems (SMS) are fully integrated, collaborative manufacturing systems that will respond in real time to meet changing demands and conditions in factories, in their supply network, and in customer needs [1] [2] [3] [4]. SMS requires the digitalization and integration of components of a manufacturing enterprise including manufacturing resources such as computer numerical control (CNC) machining systems [5].

A CNC machining system is a fundamental element in production systems and typically consists of a machine tool, cutting tools, auxiliary devices, material-handling devices, and fixtures. A CNC machine model is a conceptual representation of the machine tool and has a logical framework that enables the representation of the machine's functionalities. The information built into a machine model can be used throughout the life cycle of a machining system and by various users in the decision-making processes. Examples of model use include manufacturing capability evaluation, process validation, and production planning [6] [7]. It consists of modules for describing the configuration of the overall structure, geometric shapes of the mechanical units, as well as the kinematic relationships between the mechanical units of the machine. The kinematics model of a machine tool defines the motion constraints for machine components that are related to each other [8]. For example, a five-axis machine is generally defined by two rotational axes to rotate and tilt either the tool or the workpiece and three orthogonal linear axes x, y, and z. The machining functional properties, i.e., mechanical and kinematical properties in the machine model, will define and constrain the movements and speeds of axes [7]. Simulation of kinematics helps identify manufacturing issues at an early stage and correct them before production. Those issues could be errors in the tool path, collisions between machine components and machined parts, and poor quality of the final product. Simulation is the safest and most cost-effective way for verification of a multi-axis program, and it supports the concept of virtual machining [6] [9].

Computer-aided (CAx) tools normally provide a virtual environment that enables the simulation of machining processes with a realistic representation of the kinematics, static, and dynamic behavior of the real machine tool [6]. The x in CAx is an abbreviation for the family of computer-aided tools that are used to create virtual environments, for example, Computer Aided Design (CAD), Computer Aided Manufacturing (CAM), or Computer Aided Engineering (CAE). A variety of commercial CAx tools from different vendors are available and have been used by manufacturers to represent their products and resources to support design, operation, and maintenance activities. These activities involve process planning, tool path verification, cost estimation, process simulation, and CNC programming [6] [7].

With the multiple CAx tools provided by different vendors serving the same purpose, barriers for sharing and exchanging machine models with kinematic and geometric information between different systems exist [10]. Because each vendor has its own CAx environment, which is non-homogenous, users are stuck with the specific format of the CAx software they use. Redundant efforts have to be made for recreating the same machine model using different CAx tools within a company; machine models with complex kinematics may be difficult or very time-consuming to remodel or convert. In addition, all these issues also make it difficult to efficiently define and analyze manufacturing capabilities for production planning and equipment

Baerring, Maja; Shao, Guodong; Helu, Moneer; Johansson, Bjorn. "A Case Study for Modeling Machine Tool Systems Using Standard Representations." Presented at ITU Kaleidoscope 2020. December 07, 2020 - December 11, 2020.

procurement. For example, when purchasing a CNC machine tool, it is normally hard for final users to determine whether their workpiece(s) will fit on that machine, or even if they do fit, is there an accurate, efficient location for the parts? Having machine models beforehand will allow users to better understand the machines' capability and easily compare candidate machines through "try before they buy." However, it is impossible for users to gain access to all the vendor-specific tools and machine models before the procurement. A neutral format, which is a non-proprietary format that can be used to represent machine models and recognized by all vendors, of the machine models will provide the final users a convenient way to evaluate the capabilities of candidate machine tools.

Efforts for international standards, to govern the representation of geometrical and functional information, have been made in parallel as the machine tool manufacturers developed their proprietary techniques [7]. Attempts have been made to provide a standardized solution for model and data exchange between CAx systems, but so far it is mainly the product geometric data and definition that have been widely exchanged by the support of ISO 10303, which is also called Standard Exchange of Product Data (STEP). Many CAx tools can export geometric models to the STEP format and vice versa. Standard solutions for the exchange of kinematics information have not been used in practice and industry does need a standard way to exchange complete machine model information including the product geometry, kinematics, tolerances, and classification [10].

The increasing use of software to represent a machining system in a virtual representation, from a manufacturing perspective, implies an increased need to be able to reuse the information. Manufacturing companies are investing more to digitizing their enterprises and as more information is digitally available, the interest and urgency of information reuse will increase. Interoperability for systems and models will be more crucial and will motivate manufacturers to seek solutions that support standard representations of their resource, product, and production data. The stakeholders to this approach are manufacturing companies that need to exchange these kinds of data, both internally and externally. For example, internally, the same machine model may be developed multiple times using different software, models have issues with different version of the software, or different units used by the different component models that need to work together; externally, model exchanges and communications among supply chain partners, CAx developers, and machine vendors may be required.

This paper introduces a case study that demonstrates the feasibility of representing a complete machining model, including both geometric and kinematics information, using the STEP standard. This case study reports the initial effort of converting a vendor-specific (PTC Creo) machine model to the STEP AP 242 representation. PTC Creo was selected because (1) it is a commonly used CAx tools and (2) we have a machine model available in Creo format for this case study. The research contributions of this paper include (1) a general approach for converting a vendor-specific machine model in proprietary format to a standard format (STEP) and (2) lessons learned through the implementation of the case study.

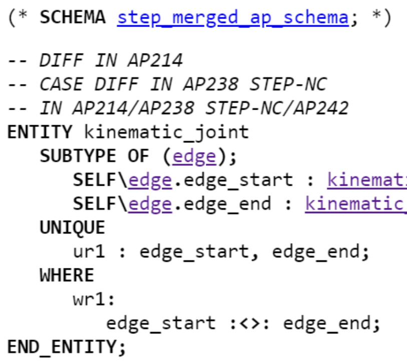
The rest of the paper is organized as follows: Section 2 discusses the existing relevant standards and related efforts for solving the problem of non-homogenous CAx environments. Section 3 presents the general approach of how to address this problem so that a specific approach can be derived depending on the CAx tools used and interfaces required. Section 4 introduces the context and settings for this specific case, and presents the development of the STEP generator. Section 5 discusses the challenges encountered in this study and finally Section 6 concludes the paper.

2. RELEVANT STANDARDS AND RELATED EFFORTS

This section discusses relevant standards that support the work, interfaces that enable the conversion of data from proprietary formats to the standard formats, and related efforts in the field.

2.1 Standards: ISO 10303 – STEP and ASME B5.59

Two standards for representing machine models are briefly introduced in this subsection: ISO 10303 and ASME B5.59. The standard ISO 10303, or STEP, was developed to unambiguously represent and exchange computer-interpretable information for a product [11]. STEP consists of a set of standards to facilitate data modeling throughout the entire lifecycle of a product, and has become widely accepted and applied internationally for exchanging product data in the manufacturing field. Information models and nearly all parts of STEP are defined using the EXPRESS modeling language, the standard ISO 10303-11 [12] [13]. The EXPRESS modeling language defines entities and the relationships between entities. Files that are created based on this standard are also referred to as physical files or part21/p21 files. The instances of an entity can be exchanged by the support of p21 files or shared within applications through the Standard Data Access Interface (SDAI) [14]. The information models can be categorized into application protocols (AP) or integrated resources (IR). APs are developed for specific application domains, such as aerospace in AP 203 and automotive in AP 214 and they are designed for fulfilling the industrial requirements [12]. AP 242, managed model-based 3D engineering, presents a data model schema to integrate the kinematics, geometry, and assembly models [15]. However, AP 242 is still a work in progress and has not been widely implemented in industry [10]. Figure 1 provides an example of the EXPRESS schema for AP 242, which defines a kinematic joint. [16] lists a complete documentation of the AP242 EXPRESS schema. IRs are context-independent, and an example of IR is ISO 10303-105 that defines an IR for kinematics data. IR for kinematics for 10303-105 specifies the structure, motion, and analysis for kinematics mechanism and is possible to use in any industrial domain [17] [18].



```
(* SCHEMA step_merged_ap_schema; *)

-- DIFF IN AP214
-- CASE DIFF IN AP238 STEP-NC
-- IN AP214/AP238 STEP-NC/AP242
ENTITY kinematic_joint
   SUBTYPE OF (edge);
      SELF\edge.edge_start : kinematic_link;
      SELF\edge.edge_end : kinematic_link;
   UNIQUE
      ur1 : edge_start, edge_end;
   WHERE
      wr1:
         edge_start :<>: edge_end;
END_ENTITY;
```

Figure 1 – An example from of the AP242 EXPRESS schema for defining kinematic joints [16]

ISO is not the only standardization organization involved in standardizing how machine models can be represented on a neutral format. Another standard that defines information models and formats for machine tool data is B5.59 from the American standardization organization, American Society of Mechanical Engineers (ASME) [9]. The eXtensible Markup Language (XML) is used for representing the specification of machine tools (milling and turning machines). The focus of the standard is on properties that describe capabilities and performance of a machine tool at a specific instance of its life cycle, e.g., in the specification or operation stage of the machine tool. The standardization efforts made from multiple standardization bodies indicate the importance of this topic and a standardized solution is needed for sharing and exchanging manufacturing resource and product models.

2.2 JSDAI

There are several available interfaces to support the translation of machine models in a vendor-specific format to the STEP format, e.g., STEP Tools [19], OpenCascade [20], PythonOCC [21], and Java-based SDAI (JSDAI) [22]. JSDAI was selected for the case study reported in this paper because it supports most of the APs in ISO 10303 and is a Java-based open source Application Programming Interface (API). JSDAI also supports the development of EXPRESS data models and their implementation in Java. It enables the reading, writing, and runtime manipulation of object-oriented data defined according to an EXPRESS data model. JSDAI provides a library that contains EXPRESS schemas for most APs in ISO 10303. JSDAI uses the EXPRESS schema defined for AP 242 to represent the kinematics information.

JSDAI facilitates the linking of CAD, CAM, CAE, CNC, Product Data Management (PDM), and Product Lifecycle Management (PLM) systems [22] [23].

2.3 Related efforts

Since the introduction of STEP in 1994, examples of how it can be used to share data on a standard format have been reported in scientific publications. Among the contributors, Li et al. [10] [24] have made efforts for converting kinematics modeling using Siemens' NX CAD software to the STEP format. The case study in this paper uses a similar approach, but a new STEP generator was developed for a different CAD tool and applied to a different machine model.

3. THE GENERAL APPROACH FOR CONVERTING MACHINE MODELS

To extract kinematics data from a machine model defined in a vendor-specific CAD system and integrate it with a STEP model that contains the geometric data of the same machine model, an application needs to be developed. Figure 2 shows the general approach of how this could be done as a guideline for readers with various CAx tool to follow. Section 4 will explain the case specific settings for the case study in this paper and Figure 3 depicts the specific procedure for the case study.

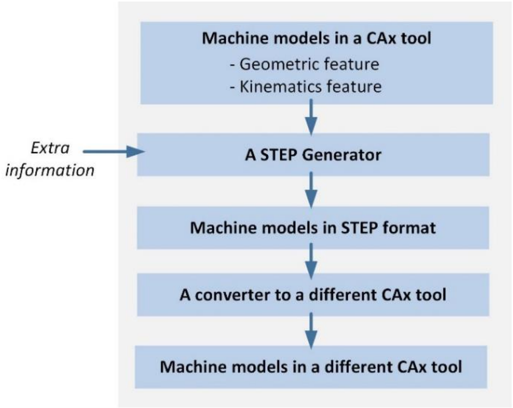

Machine models in a CAx tool
- Geometric feature
- Kinematics feature
Extra information
A STEP Generator
Machine models in STEP format
A converter to a different CAx tool
Machine models in a different CAx tool

Figure 2 – The general approach for creating a complete machine model in STEP format with both geometric and kinematics information

A machine model may be developed in a vendor-specific format with a complete description including geometric and kinematics information; this is a foundation for us to be able to translate the complete virtual machine model to a standard format. Most CAx software today provides the functionality to automatically export geometric information in the STEP format. However, a vendor-specific application (i.e., an interface or adapter) is required to extract the kinematics data. Examples of interfaces that support the development of such applications include J-Link for PTC Creo and NX Open for Siemens. These interfaces of vendor-specific tools enable the development of the STEP generator.

After both geometric and kinematics data sets have been extracted from the machine model in a CAx tool, the "STEP Generator" integrates them into a complete machine model in STEP. For example, JSDAI can be applied for integrating the information from both sources to create a final complete STEP model containing all the information of the machine model from the CAx tool. To ensure that the STEP file is complete, there may be extra information such as users' input that needs to be added. With vendor-specific converters/adapters, the complete machine model in STEP format can be imported into other CAx environments. The vendor-specific STEP generators vary depending on the specific interface requirements and programming language used. The STEP generator used for the case presented in this paper will be further explained in the next section.

4. A CASE STUDY – A STEP GENERATOR FOR PTC CREO

We have applied the approach described in Section 3 to a specific use case. Figure 3 shows an instance of the general approach depicted in Figure 2. The machine tool model is defined in PTC's CAD software, Creo. The geometric information of the model is exported to a STEP file. The STEP file will be integrated with the kinematics information generated by using J-Link and JSDAI. J-Link is a Java-based API that is provided by PTC to enable the interactions between the machine model in Creo and other applications. Through J-link, kinematics information from the model can be extracted. The geometric information in the STEP file and the kinematics data are integrated into one STEP file by the STEP generator. The following subsections will explain each step in more details.

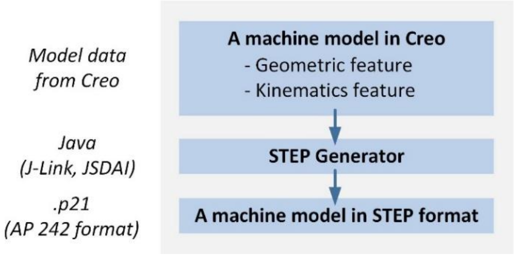


Figure 3 – The case study approach: integrating a Creo machine model's kinematics and geometric information in STEP AP242

4.1 Machine model in Creo

The machine model in the case study is a 5-axis Hurco CNC machine tool, VM10UI. It is developed in PTC's CAD environment, Creo Parametric professional version 6. The J-Link API is an add-on module of the software. The Hurco machine model in Creo is shown in Figure 4.

It consists of a spindle and the y slide representing a machining table. The table can move in x-axis and y-axis, and rotate to adjust the angle of a part in relation to the spindle head. VM10UI_MAINCYS is the coordination system for the machine model of the HURCO machine, and ADTM1, ADTM2, ADTM3, and ADMT4 are the four planes that serve as reference for a planar surface. Defined planar enables motion settings of the machine parts with the same reference. The tree structure to the left in Figure 4 contains the kinematics information on how the different parts and assemblies of the machine model are related to each other, which determines how they move. HURCO_VM10UI is the name and model of the machine tool, and FRAME, Y_SLIDE and SPINDLE_HEAD are components constituting the machine model. Each of the components in the tree structure has a breakdown structure where more information is contained for the machine model, such as the kinematic information, which defines either the rotational or translational movement in the x- and y-axis. Figure 5 shows the breakdown structure for the component SPDINLE_HEAD.

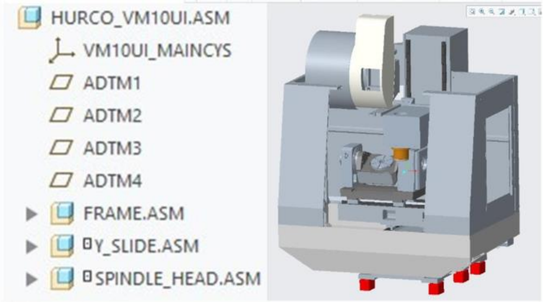


Figure 4 – The Hurco machine tool model in Creo Parametric

A representation of the spindle head of the machine (marked with green lines) is shown in Figure 5. The tree structure of the spindle head is expanded in the list to the left, representing the information for the spindle head, which includes kinematics information about the alignment and rotation. Placement contains the kinematic data with information of the axis alignment and the rotation of the spindle head. A tool is attached in the spindle head and the rotational movement determines how material is removed during machining. DEFAULT_CSYS is the coordination system defined for the spindle head and ASM_RIGHT, ASM_TOP, and ASM_FRONT are planes applied for this machine part.

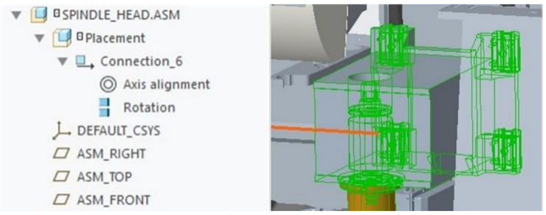
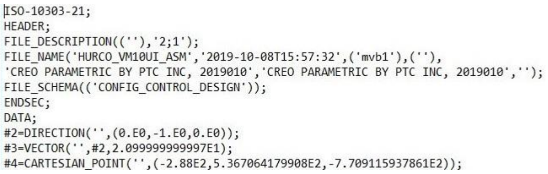


Figure 5 – A view of the spindle of the machine and its tree structure

4.2 Geometric data in a STEP file

The geometric information of the machine model is exported into a STEP file using the standard interface provided by Creo. The export functionality automatically generates a .stp file; a portion of the exported STEP file is shown in Figure 6. The STEP file (.stp) contains all the geometric data starting from the line defining DATA. The schema used for defining the model is CONFIG_CONTROL_DESIGN from AP203.



```
ISO-10303-21;
HEADER;
FILE_DESCRIPTION((''),'2;1');
FILE_NAME('HURCO_VM10UI_ASM','2019-10-08T15:57:32',('mvb1'),(''),
'CREO PARAMETRIC BY PTC INC, 2019010','CREO PARAMETRIC BY PTC INC, 2019010','');
FILE_SCHEMA(('CONFIG_CONTROL_DESIGN'));
ENDSEC;
DATA;
#2=DIRECTION('',(0.E0,-1.E0,0.E0));
#3=VECTOR('',#2,2.099999999997E1);
#4=CARTESIAN_POINT('',(-2.88E2,5.367064179908E2,-7.709115937861E2));
```

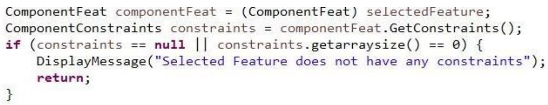
Figure 6 – A snap view of the STEP file that is exported from Creo and contains all the geometric information of the machine tool model

4.3 The STEP Generator using JSDAI and J-Link

The Java development environment used in this study is Eclipse. JSDAI provides plug-ins that are compatible with Eclipse. JSDAI also provides an EXPRESS compiler for compiling the EXPRESS files and creating .jar files for use in Java programs to represent the data model.

The STEP Generator uses an iterative process to evaluate the characteristics of the kinematics information and add it accordingly to the STEP file. To allow JSDAI to manipulate the model data, the read-and-write access is used for accessing the data in the generated STEP file with geometric data and for writing kinematics data to the STEP file to create a complete STEP model according to the AP 242 EXPRESS model. A repository is created for JSDAI to store the temporary kinematics data.

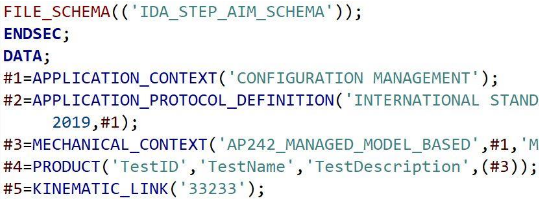
The kinematics information in the Hurco machine model is defined as constraints. For each constraint, an array will be created to store the data. This data is written to the STEP model according to the EXPRESS schema used, i.e., AP 242, and an example is shown in Figure 7.



```
ComponentFeat componentFeat = (ComponentFeat) selectedFeature;
ComponentConstraints constraints = componentFeat.GetConstraints();
if (constraints == null || constraints.getarraysize() == 0) {
    DisplayMessage("Selected Feature does not have any constraints");
    return;
}
```

Figure 7 – An example of code where the kinematic data of model feature is extracted for conversion to the STEP format

After going through all the kinematics constraints, a STEP model representing both geometrical and kinematics data is generated from JSDAI as a .stp-file and a section of such a model is shown in Figure 8.



```
FILE_SCHEMA(('IDA_STEP_AIM_SCHEMA'));
ENDSEC;
DATA;
#1=APPLICATION_CONTEXT('CONFIGURATION MANAGEMENT');
#2=APPLICATION_PROTOCOL_DEFINITION('INTERNATIONAL STAND
     2019,#1);
#3=MECHANICAL_CONTEXT('AP242_MANAGED_MODEL_BASED',#1,'M
#4=PRODUCT('TestID','TestName','TestDescription',(#3));
#5=KINEMATIC_LINK('33233');
```

Figure 8 – An example of a complete STEP file (.stp file) generated by the STEP generator

5. DISCUSSION

This work contributes to the field of system interoperability and information reuse for machine modeling. During this study, challenges and issues have been identified, and more research and development efforts are required to address them. The challenges are elaborated in the following subsections from different perspectives: (1) challenges with the applications of STEP, JSDAI, and J-link, (2) challenges with converting the machine models and kinematics information including aspects of verification and validation of the developed approach, and (3) challenges with the commercial software and data reuse for end users.

5.1 Challenges with the applications of STEP, JSDAI, and J-Link

The STEP standard has been a work in progress since its introduction in 1994 and there are continuous improvements and new additions to it. One of the latest developments is the AP 242 edition 2 that integrates the definitions from both AP 203 and AP 214, which are originally developed for different manufacturing industries. The new AP becomes more complex and is harder for users to understand and use. Because the STEP definitions are cumbersome, it requires a specific software for editing and manipulating a STEP file.

JSDAI covers most definitions that are needed for writing, reading, and modifying STEP models. This makes JSDAI applicable to the development of the kind of STEP generators we described in this paper. However, the complication of the STEP definitions has also added more complexity to the JSDAI applications. Since JSDAI is an open source API, there are few examples demonstrating real use cases of where JSDAI has been used. The technical support from the developer of JSDAI is hard to get and the documentation of JSDAI is not up to date. With better documentation, more examples, and further developments, JSDAI can facilitate the implementation of the STEP standard more efficiently. The effort required for this implementation was about 4 months for a person with basic programming skills. By referring the approach proposed and the lessons learned in this paper, an industry application could be implemented in a shorter time. More JSDAI implementations would also motivate the enhancement and the support of technology.

J-Link enables the interaction between a Creo model and JSDAI. J-Link provides documentation, guidelines, and program examples to support developers and users of Creo. Since the interface and programming environment are vendor-specific, developers of STEP generators will need to have knowledge and programming skills for multiple tools. How the machine model is defined in a CAx specific software will impact how the data that represents the feature, part, and object of the machine can be manipulated.

5.2 Challenges with the converting of machine models and kinematics information

In this study, a couple of constraints (kinematic properties/pairs) have been converted to the STEP format. However, automatically identifying and converting all constraints of the machine model is still challenging. More effort is needed to ensure the correct usage of the EXPRESS schema when generating the STEP representations automatically, i.e., the machine model data exported from the CAD software is converted correctly to the STEP format. This involve the STEP generator including the integration of J-Link and JSDAI applications for exporting the complete machine model on a neutral format. This leads to an important topic, the verification and validation (V&V) of the converted machine model. Is there anything missing during the model conversion? Does the newly generated STEP machine model exactly represent the original vendor-specific machine model? Although V&V has not been a focus of this study, techniques for V&V of the machine models have been investigated.

Kinematics information is crucial for the behavior of a machining system and the accuracy of the kinematics model determines the precision of the overall machining. Kinematics modeling is one of the most common sources of errors for a machine model. Therefore, when remodeling of kinematics information of a machine tool, it is important to ensure the kinematics information is converted completely and correctly between various systems and formats using the STEP generators. In order to do that, a fundamental requirement is that the coordinate systems in ISO 10303-105 (STEP part21 file) and in the CAD software need to be the same. In ISO 10303-105, a link frame is used to define the local coordinate system of a kinematic pair and all relevant geometric definitions are defined relative to this link frame. On the other hand, commercial CAD software has its own way of defining coordinate systems. Many of them use a world coordinate system (or a global coordinate system). So before converting the machine model to the STEP format, the coordinate system needs to be converted; this includes the location and orientation information of each pair, in Creo, and it is for each constraint. The terminology usage in different CAD environments for the same concept also causes a lot of confusion which was encountered in the case presented here; what is referred to as a part in one software may be called a feature in another software. What is called constraints in Creo will be called pair or link in STEP. This poses implications for the extraction of kinematics data and needs to be adjusted for each vendor-specific CAD software. Since the terminology for each vendor-specific software already exists and is being used, this needs to be considered during the model conversion. It could also be argued that the terminology should be standardized but it will be a long way to go not only for the development of the standard, but also for all vendors to comply with the standard. Note that also the complexity of the conversion will increase with the number of axes, e.g., a five-axis machine is more complex than a three-axis machine.

There are also remaining challenges for the definition of kinematics in AP 242. It was first introduced in 2014 but is still not widely used or implemented in industry. Is AP 242 a perfect solution for representing all machine models? In other words, are the definitions in AP 242 complete for all the needs of the kinematics definitions? This needs to be further investigated.

The current situation is that kinematics information is managed manually in some companies by using a text-based description or a spreadsheet-like tool. Manual steps involving humans always has the risk of creating errors. Most other smaller companies are not even capable of dealing with kinematics settings at all because of the lack of knowledge and access to the information. This may cause production delay, product quality issues, more vendor dependency, and cost increase.

5.3 Challenges with commercial software and data reuse for end users

Even though STEP is now an integrated interface in most CAD software, there is still an unwillingness from the CAx developers for further implementation of STEP representations of kinematics modeling. This is because of the complexity of the implementation, but also because the vendors and solution providers would like to take advantage of the situation with customer retentions and lock-in effects [7]. Most CAx software developers provide vendor-specific solutions so that customers need to depend on their software. This situation causes information silos and makes it difficult for interoperability. It also causes more issues for model and data reuse for manufacturers because of the diverse landscape of software and systems they use. It used to be the same situations for post-processing and Geometric Dimensioning and Tolerancing (GD&T), which have currently been implemented using a standard format by most CAx vendors. So, we hope kinematics modeling is the next one that people will turn to standardized solutions since there is a clear need for it from the manufacturing community that could motivate the CAx software to provide a standardized solution in the same way as it is for GD&T now.

6. CONCLUSIONS AND FUTURE WORK

The case study presented in this paper demonstrates the feasibility of generating a complete STEP model that includes both geometric and kinematics information of a machine model. This is done by developing a STEP generator to extract geometric and kinematics data and integrate it into the STEP model according to the EXPRESS schema. The results of the paper include:

- A general approach has been developed for how kinematic and geometrical data can be extracted to a neutral format. The general approach is meant to be applicable to all CAx tools.
- Explaining how the general approach was used for a case specific setting, including a description of the interfaces and software that were used. This is specific for the tools selected and interfaces determined for the case study.
- The translation of a machine model to the STEP standard format was explained for a real machine model developed in the CAD software, PTC Creo.
- The case study with the PTC Creo machine model serves as a feasibility study and demonstrates step-by-step how this can be done.

The work has real industrial impact and the standards-based digital representations of a complete machine tool model enables better information reuse, better interoperability, and more consistent management. It will help support decision making throughout the different phases of a production system including machine tool procurement, efficient machining capability definition and analysis, dynamic planning and scheduling by facilitating last-minute adjustments to adapt current conditions, and configuration validation. Furthermore, it will save both time and money for the manufacturing companies.

This is a preliminary study. More real-world industrial cases will be implemented. Also, more CAx software specific adapters need to be developed. One scenario could be a real-world study with supply chains involving several companies using different systems and those companies representing both large enterprises and small and medium-sized enterprises. Supply chains in both process and discrete manufacturing may be used to demonstrate how existing challenges in information sharing and model exchange could be addressed.

Acknowledgements

The work presented in this paper has been supported by the NIST Model-Based Enterprise (MBE) program and Area of Advance Production at Chalmers University of Technology as well as the research projects Digitala Stambanan and 4S that are funded by the Swedish Agency for Innovation Systems (VINNOVA).

The authors also would like to thank Dr. Parag Vichare from University of the West of Scotland for providing the machine model, Steven Mallia from Purdue University and Yujiang Li from Siemens for their support, and colleagues from ISO TC 184 SC4 WG 15 Digital Manufacturing Implementation Forum for valuable discussions and support.

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

[1] NIST, “Product Definitions for Smart Manufacturing,” www.nist.gov. https://www.nist.gov/programs-projects/product-definitions-smart-manufacturing (accessed July 7, 2020).

[2] S.S. Shipp, N. Gupta, B. Lal, J.A. Scott, C.L. Weber, M.S. Finnin, M. Blake, S. Newsome, and S. Thomas, “Emerging global trends in advanced manufacturing,” Institute for Defense Analyses, Alexandria VA, March 2012.

[3] J. Davis, T. Edgar, J. Porter, J. Bernaden, and M. Sarli, “Smart manufacturing, manufacturing intelligence and demand-dynamic performance,” Computers & Chemical Engineering, vol 47, pp.145-56, 2012.

[4] SMLC, “Implementing 21st century smart manufacturing,” In Workshop summary report SMLC, June 2011.

[5] Y. Lu, K.C. Morris, and S. Frechette, “Current standards landscape for smart manufacturing systems,” National Institute of Standards and Technology, NISTIR 8107, 2016.

[6] P. Vichare, X. Zhang, V. Dhokia, W.M. Cheung, W. Xiao, and L. Zheng, “Computer numerical control machine tool information reusability within virtual machining systems,” Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, vol. 232 no. 4, pp. 593-604, 2018.

[7] P. Vichare, A. Nassehi, and S. Newman, “A unified manufacturing resource model for representation of computerized numerically controlled machine tools,” Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, vol. 223 no.5, pp. 463-83, 2009.

[8] T. Kjellberg, A. von Euler-Chelpin, M. Hedlind, M. Lundgren, G. Sivard, and D. Chen D, “The machine tool model – A core part of the digital factory,” CIRP Annals, vol. 58 no. 1, pp. 425-428, 2009.

[9] S. Zivanovic, N. Slavkovic, B. Kokotovic, and D. Milutinovic, “Machining Simulations of Virtual Reconfigurable 5-axis Machine Tool,” Annals of the Faculty of Engineering Hunedoara, vol. 15 no. 2, pp. 189-94, 2017.

[10] Y. Li, M. Hedlind, T. Kjellberg, and G. Sivard, ”System integration for kinematic data exchange,” International Journal of Computer Integrated Manufacturing, vol. 28 no. 1, pp. 87-97, 2015.

[11] ISO 10303-1, Industrial Automation Systems and Integration–Product Data Representation and Exchange–Part 1: Overview and Fundamental Principles, 1994.

[12] T. Kramer and X. Xu, “STEP in a nutshell,” In Advanced design and manufacturing based on STEP, Springer, London, pp. 1-22, 2009.

[13] D. Loffredo, “Fundamentals of STEP implementation,” STEP Tools Inc., pp. 1-12, 1999.

[14] ISO 10303-22: Industrial automation systems and integration-Product data representation and exchange-Part 22: Implementation methods: Standard data access interface specification, 1998.

[15] STEP AP242 Project, http://www.ap242.org/other-related-standards (accessed July 7, 2020).

[16] STEP Tools, STEP AIM, https://www.steptools.com/stds/stp_aim/html/ (accessed July 7, 2020).

[17] S.J. Kemmerer, “STEP: the grand experience,” Washington D.C., US, G.P.O., October 1999.

[18] ISO 10303-105: Industrial Automation Systems and Integration-Product Data Representation and Exchange—Part 105: Integrated Application Resource: Kinematics. International Organization for Standardization, 1996.

[19] STEP Tools, STEP ARM, http://www.steptools.com/stds/stp_expg/arm.html (accessed July 7, 2020).

[20] OpenCascade, https://www.opencascade.com/doc/occt-7.0.0/overview/html/occt_user_guides__step.html (accessed July 7, 2020).

[21] PythonOCC, http://www.pythonocc.org/category/features_overview/ (accessed July 7, 2020).

[22] JSDAI. Homepage for JSDAI by LKSoft. http://www.jsdai.net/ (accessed on January 15, 2020).

[23] C. Pan, S.S. Smith, and G.C. Smith, “Automatic assembly sequence planning from STEP CAD files,” International Journal of Computer Integrated Manufacturing, vol. 19 no. 8, pp. 775-83, 2006.

[24] Y. Li, “Implementation and evaluation of kinematics mechanism modeling based on ISO 10303 STEP,” Master Thesis, Royal Institute of Technology, Stockholm, Sweden, 2011.

Semi-Automated Analysis of Large Privacy Policy Corpora

Alden Dima
National Institute of Standards and Technology
alden.dima@nist.gov

Aaron Massey
University of Maryland Baltimore County
akmassey@umbc.edu

Abstract

Regulators, policy makers, and consumers are interested in proactively identifying services with acceptable or compliant data use policies, privacy policies, and terms of service. Academic requirements engineering researchers and legal scholars have developed qualitative, manual approaches to conducting requirements analysis of policy documents to identify concerns and compare services against preferences or standards. In this research, we develop and present an approach to conducting large-scale, qualitative, prospective analyses of policy documents with respect to the wide-variety of normative concerns found in policy documents. Our approach uses techniques from natural language processing, including topic modeling and summarization. We evaluate our approach in an exploratory case study that attempts to replicate a manual legal analysis of roughly 200 privacy policies from seven domains in a semi-automated fashion at a larger scale. Our findings suggest that this approach is promising for some concerns.

1. Introduction

Privacy policies support commerce by informing potential customers of business practices involving their data. The U.S. Federal Trade Commission (FTC) investigates incongruities between stated and actual business practices. Mishandling of privacy can be costly as two recent cases highlight. In July 2019, Equifax agreed to pay at least $650 million for a data breach that affected some 150 million people [1]. That month, Facebook was fined $5 billion for violating customer privacy and was forced to revamp its privacy protection practices [2].

These cases highlight the reactive nature of investigations, which typically begin only after a complaint is received. There is a need for rapid analysis of many privacy policy documents from websites across multiple industries [3]. Unfortunately, proactively evaluating privacy practices is tedious and challenging. These analyses often require significant manual effort. For example, in 2016, Marotta-Wurgler (Section 3) published a study of 261 privacy policies whose goal was to measure online site compliance with self-regulatory guidelines [4]. A team of nine investigators hand coded the presence of 49 recommended practices [4]. With over one billion web servers [5], manual analyses cannot scale to the Web.

We propose a semi-automated method based on the natural language processing of a large privacy policy corpus that relies on a manual analysis of a smaller subset. We wish to determine the extent that we can automate large-scale analyses using a sentence similarity baseline method. Using a corpus of two thousand documents, we will focus on the notice-related concerns from the Marotta-Wurgler study. Our research questions are:

RQ1: Which of the Marotta-Wurgler notice-related privacy concerns can be identified from our corpus using a sentence similarity-based approach?

RQ2: How well does our similarity-based approach identify privacy concerns in our policy corpus?

Our results indicate that semi-automated analysis can indeed allow for future work to target much larger corpora. We were able to identify relevant sentences containing keywords for 9 of the 21 notice-related concerns identified by Marotta-Wurgler with an average sensitivity and specificity of 93.8 % and 98.3 %, respectively. When combined with other concerns whose sentences were identified with higher average specificity (97.8 %.) but lower average sensitivity (35.6 %), sentences from 17 of the 21 notice-related privacy concerns were identified. This suggests that, while not all relevant sentences can be found, most of those found will be.

Our work proceeds as follows. In Section 2, we survey concepts and techniques necessary for this work. Section 3 describes two prior analyses of privacy policy corpora. In Sections 4, and 5, respectively, we describe our methodology and evaluation process, and present the results of evaluating our method with a corpus of 2061 policies. We continue with a discussion our results and the limitations of our approach in Sections 6 and 7, respectively, and conclude with a summary of our findings in Section 8.

Dima, Alden; Massey, Aaron. "Semi-Automated Analysis of Large Privacy Policy Corpora." Presented at 54th Hawaii International Conference on System Sciences (HICSS 2021), Kauai, HI, US. January 05, 2021 - January 08, 2021.

2. Background

In this section, we describe techniques and concepts that are related to this study. We begin by discussing extractive text summarization, followed by vector space models. We then discuss semantic similarities and conclude with the total recall problem.

Extractive Text Summarization Our approach depends on the identification of sentences similar to manually-selected exemplar sentences relevant to a privacy concern. We use extractive text summarization to find candidate exemplar sentences likely to match the greatest number of relevant sentences with the same keywords. A few of the highest-ranking sentences can serve as a summary. We chose LexRank, an extractive text summarization algorithm that uses similarity graphs to connect sentences whose cosine similarities exceed a threshold [6].

Vector Space Models In order to identify similar sentences, we must first find a suitable computer representation to encode their features. Text inherently lends itself to sparse representations requiring large amounts of computer memory and computing time [7, 8]. For some of our work, we addressed this via a dimensionality reduction technique, latent semantic indexing (LSI), which uses a singular value decomposition to project the words into a smaller concept space [9, 10, 11, 12]. As an added benefit, Words that appear in similar contexts are projected into the same concepts so that documents can be close to each other without sharing common words [11].

Newer techniques such as word2vec [13], GloVe [14], Elmo [15], and BERT [16] use neural networks to achieve dense representations. These techniques are more sensitive to the surrounding word context and give additional semantic information. We chose a preexisting BERT embedding for a portion of this work and compared the results obtained with those from using LSI.

Semantic Similarity Once we identified the exemplar sentences and calculated the sentence features, we needed a means to identify sentences that are similar to the exemplars. We used cosine similarity, which treats the word features as vectors and uses the angle between two vectors; the greater the angle, the less similar the words [17].

The Total Recall Problem The semi-automated analysis of privacy policies shares aspects with the Total Recall Problem (TPC), an information retrieval problem where only a small fraction of a population is relevant and each member of that population can be inspected [18, 19]. The goal is to reduce the cost of high recall with a human in the loop [18, 19]. The best strategy for language-based TPC is to apply active learning with vector-based features to a manually-labeled subset and then use the trained system to prioritize the remaining items [19]. A challenge is knowing when to stop the iterative process of manually labeling positive examples and use the trained model to predict the unlabeled ones [19].

Our baseline method is similar to the active learning approach. For example, we use vector-based features and identify positive examples (the exemplar sentences). However, our method does not share the active learning's iterative nature. We evaluate the resulting similarity-based classifiers on a subset of the data, but do not iteratively train them. The active learning approach to the Total Recall Problem may serve as a useful strategy for situations where we are unable to identify sentences with high sensitivity using our method.

3. Related Work

In this section, we will survey four prior studies as well a deep-learning based analysis framework.

RE 2013 Study and Corpus To allow us to generalize our results, our exploration required a sufficiently large and diverse corpus. We chose a corpus of 2061 privacy policy and related documents drawn from prior privacy document analysis work, the Google top 1000 websites, and the 2012 Fortune 500 companies [3]. It was created to determine if automated text mining can help requirements engineers determine whether a policy document contains requirements expressed as either privacy protections or vulnerabilities. Massey et al. used topic modeling to identify documents addressing concerns expressed via goals-based requirements engineering and demonstrated the ability to limit searches of the entire corpus to less than 100 documents for certain goal keywords [3].

Marotta-Wurgler Study We benefited from the privacy concerns identified in Marotta-Wurgler's 2016 study [4] of 261 privacy policies taken from seven online markets [4]. She found that the average policy only complies with 39 % of the 2012 FTC guidelines [4]. The policies were often silent in required or important areas. This causes problems because there are often no default rules to address these areas. Many companies collect information and consumers have no way of knowing how much is collected, how it is used, how long it is kept, or with whom it is shared. The guidelines do not seem to be driving policy development and the idea of companies "competing on privacy" appears to be flawed [4].

Polisis The challenges of analyzing privacy policies at scale led Harkous et al. to create Polisys, deep learning based policy query answering framework that uses convolutional neural networks classifiers, a privacy taxonomy, and a custom language model created from 130,000 privacy policies for state-of-the-art results for structured and free-form querying of privacy policies [20]. Polisys relies on an embedded privacy taxonomy, which can limit its use. Harkous et al. suggest that this can be mitigated via additional training using another annotated data set.

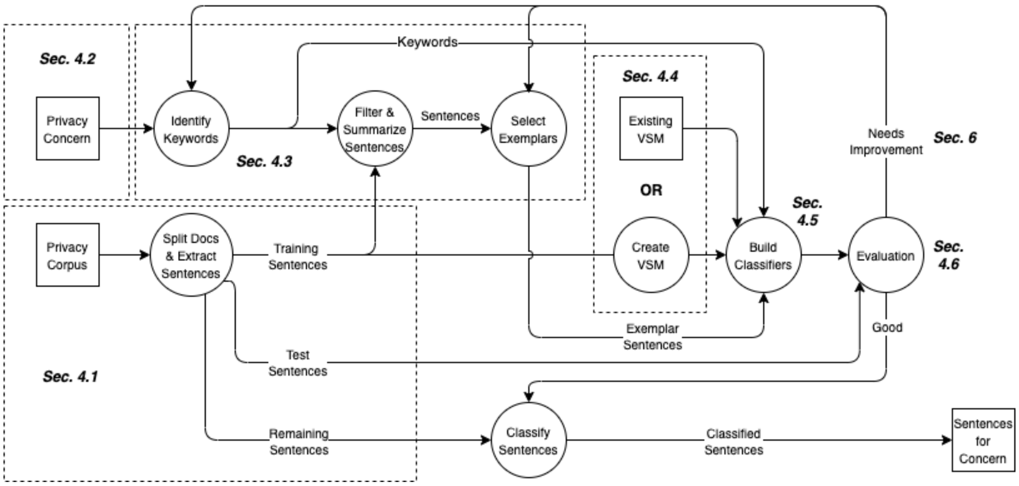


Figure 1: Data flow diagram of the semi-automated methodology for identifying sentences related to a privacy concern described in Section 4. Portions of the diagram are labeled with the section or subsection which describes them.

We believe that the method described in Section 4 can be used as part of a process to create such a data set.

GDPR Impact Studies Two recent studies attempted to ascertain the impact of the European Union's (EU) General Data Protection Regulation (GDPR) on privacy policies. The first by Degeling et al., used a combination of automated and manual methods to routinely inspect some 6,579 websites across the EU for evidence of changes due to the enactment of the GDPR [21]. They identified and downloaded some 112,041 privacy policies into a database, extracted the individual sentences, identified changes with hash functions and analyzed sentences for occurrences of GDPR-related phrases. They found that most EU websites had made changes to address the GDPR, though not all new requirements were being met.

Linden et al. also addressed the impact of the GDPR on privacy policies. They collected 6,278 pairs of pre- and post-GDPR versions of English-language privacy policies and examined them for changes in five areas: visual presentation, syntactic text features, category-based coverage, compliance to a small set of requirements, and the specificity of certain privacy practices [22]. They identified the English-language privacy policies using a convolutional neural network and made use of Polisis [20] to automatically label text segments. They conclude that the GDPR has driven many recent changes in privacy policies, particularly in the EU [22].

4. Methodology

Our ultimate goal is to enable the future analysis of large Web-based privacy policy corpora to determine the extent to which specific privacy concerns are being addressed. While previous studies have used manual methods on smaller corpora, they will be intractable for our future intended corpora. We plan to leverage a semi-automated analysis of a small subset of a large corpus to perform an automated analysis of the rest. Our approach requires dividing a large corpus into two portions: a smaller subset of the privacy policies amenable to combined manual and automated analyses and the much larger remainder which can only be analyzed using automated methods. Our goal for this work was to explore whether this approach is feasible and to identify potential issues for future research. The methodology described below was intended to represent a minimally viable technique and serves to establish a baseline for more refined approaches. We used the following criteria to answer our research questions:

RQ1 Measures: We determined which of the notice-related privacy concerns could be identified using our similarity-based approach while

(a) *Excluding* privacy concerns that address external language or entities outside of a privacy policy.

(b) *Excluding* privacy concerns that have no identifiable keywords and require human interpretation to determine whether they have been addressed.

(c) *Excluding* privacy concerns which do not have both exemplar sentences containing relevant keywords and higher sensitivity and specificity similarity-based classification results with our test corpus.

RQ2 Measures: We determined how well our similarity-based approach identified notice-related concerns by clustering the classification results and identifying the different cases as described in Section 4.6.

Table 1: Notice-related concerns from Marotta-Wurgler's study. Privacy concerns with a † are those for which we are able to identify sentences using our method (Section 5).

Concern	Description	Concern	Description
N1	Policy is accessible through direct link from the homepage	N12	PII used internally for business purposes
N2	Users asked for consent when signing up via clickwrap	N13	PII used for stated, context-specific purposes
N3	Layered or short notice is collected and stored	N14 †	Profile, picture, or other data may be used in ads
N4	Contact data is collected and stored	N15 †	Third party may place ads that track user behavior
N5 †	Computer data is collected and stored	N16	Recipients of shared or sold data are identified
N6 †	Interactive data is collected and stored	N17	Words such as "affiliates" are defined, if used
N7 †	Financial information is collected and stored	N18 †	Company alerts user to material changes in policy
N8	Content is collected and stored	N19 †	User must explicitly assent to material changes
N9 †	Sensitive information is collected and stored	N20 †	Material changes are retroactive
N10	Geolocation information is collected and stored	N21	Describes data procedures if company is sold or closes
N11	Cookies used		

We quantified the overall extent of these cases with the fraction of the test corpus sentences containing relevant keywords associated with the results for each case.

Our methodology, which is illustrated in Fig. 1 and described below in Sections 4.1 through 4.6, began with the selection of the small corpus and its division into training and test sets (Section 4.1). We then chose the privacy concerns to study (Section 4.2), identified their exemplar sentences from the training set (Section 4.3), created a vector space model from the training set sentences (Section 4.4), and built similarity-based sentence classifiers (Section 4.5). We ended with the evaluation of the classifiers against the test set in Section 4.6.

4.1. Creation of Training and Test Sets

We began with a corpus consisting of several thousand privacy policies, the RE 2013 corpus discussed in Section 3. For the purposes of this work, this corpus will serve in the place of a smaller subset of a much larger corpus. It has been the subject of prior analyses and is well understood. We randomly divided this corpus into training and test sets using a 66/33 % split. We continued by segmenting each document of the training and test sets into individual sentences. We treated these sentences as individual documents when we searched the corpus for similar sentences to build our classifiers. These steps are depicted in the lower left portion of Fig. 1.

4.2. Identification of Privacy Concerns

The upper left corner of Fig. 1 depicts privacy concern identification. We focused on the set of 21 notice-related concerns from Marotta-Wurgler [4] listed in Table 1. The first three describe the online prominence of the policy (N1, N2, and N3). Subsequent concerns focus on the user data collected and stored (N4 through N10), the use of cookies (N11), how personal and personally identifiable information is used (N12 through N14), third parties (N15 through N17), the handling of material changes (N18 through N20), and data procedures if the site is sold or ceases to exist (N21).

Notice is central to the "notice and choice" privacy model espoused by the FTC which encourages the creation of privacy policies focusing on data collection [4]. These concerns also form the largest subset of the 49 in Marotta-Wurgler's study and we believed that their concrete nature would make them amenable to our approach.

4.3. Identification of Exemplar Sentences

In this and the following two subsections (Sections 4.4 and 4.5), we will use the privacy concern N19 ("User must explicitly assent to material changes") as the basis of a running example to explain our methodology.

We began by identifying keywords for each privacy concern (Fig. 1, upper right). For privacy concern N19, we chose "you", "opt-in", and "change" as one set of possible keywords. These seemed reasonable based on the results of text searches of the privacy corpus for "material changes."

We then filtered the training set using the stemmed keywords. The N19 keywords and sentences in the training set were stemmed. For example, the sentence describing N19:

> *You will be given the choice at that time to "opt-in" for any additional uses or disclosures of your personally identifying information and/or health-related personal information that you made available to us prior to the change in the Privacy Policy.*

became

> *you will be given the choic at that time to opt in for ani addit use or disclosur of your person identifi inform and/or health relat person inform that you made avail to us prior to the chang in the privaci polici*

Stemming simplifies text searches by normalizing words so that their different forms are made identical.

Because many sentences matched the keywords, we summarized the filtered sentences using LexRank (Section 2) to identify a few of the most relevant ones. We then selected those, which in our judgement are indicative of the privacy concern to be our exemplar sentences and excluded those that are not. For example, the previous sentence was chosen as an exemplar. Despite containing

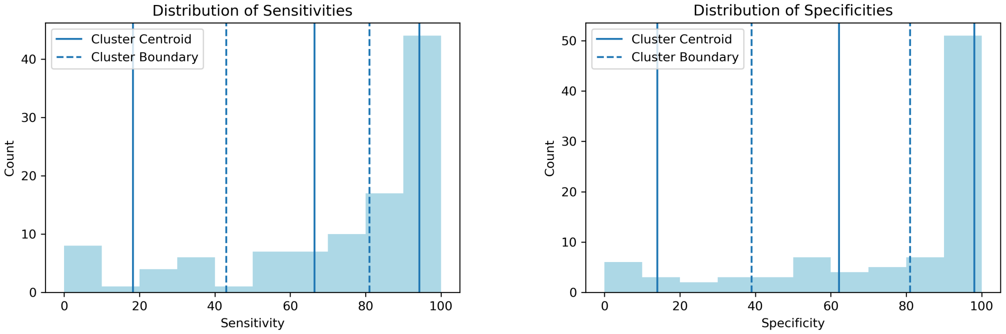


Figure 2: The distribution of sensitivities and specificities for the classifiers trained and evaluated using the process described in Sections 4.1 through 4.6. Undefined values are not shown. The solid lines denote 3-means cluster centroids and the dashed lines represent the boundaries between the clusters.

the keywords, the following sentence is not concerned with material changes and was not chosen:

> *You may change your interests at any time and may opt-in or opt-out of any marketing / promotional / newsletters mailings.*

4.4. Creation of Vector Space Models

We evaluated two approaches for creating vector-space models (VSM) of our privacy corpus. The first approach used LSI and the second a pre-existing BERT embedding (Fig 1, middle right).

For the LSI-based approach, we first created a traditional vector-space model (VSM) of the training set sentences by stemming their words. Unlike typical practice, we did not remove stop words because they help determine if a sentence should be selected. We then created bag of words (BOW) for each sentence and applied a term frequency-inverse document frequency (TFIDF) transformation followed by latent semantic indexing (LSI) using 400 dimensions, the value determined empirically by Bradford as giving the best results for large corpora [23].

For the BERT-based approach, we used a preexisting embedding which directly embeds each sentence into a 1024-dimension space [16]. Unlike LSI, we did not first stem the words of the sentences when using this preexisting embedding as it was not created using stemmed words.

4.5. Building Sentence Classifiers

Using the exemplar sentences and VSM from Sections 4.3 and 4.4, we created both LSI- and BERT-based sentence classifiers for each concern (Fig 1, right).

These classifiers begin by first selecting sentences from the test data that contain the predetermined keywords associated with each privacy concern and then selected those whose cosine similarity exceeds a selection threshold relative to at least one of their exemplar sentences. We chose a threshold value of 0.5 based on the intuition that for values above this threshold, the two vectors representing the sentences are more generally aligned with each other.

For example, consider a classifier that uses the exemplar sentence described in Section 4.3. It begins by using the keywords associated with that exemplar sentence, “you”, “opt-in”, “change” to filter test sentences. Let us assume that the following sentence is to be classified:

> *You will be notified if any of the material changes that affect the use of your personal information and asked to opt-in to the new use of your personal information.*

This sentence contains the keywords so the classifier will then calculate its embedding. This value is compared with that of the exemplar and the test sentence is rejected as being below the selection threshold. It may yet be classified as relevant if the classifier has another exemplar for which the sentence exceeds the similarity threshold.

For the BERT-based version of this classifier, the similarity between the exemplar exceeds the selection threshold, and the sentence is classified as relevant. The differences between the LSI- and BERT-based classifiers are presented in Section 5 and discussed in Section 6.

We chose this classification scheme to address class imbalances in the training set, to simplify classifier training, and to minimize the need for human-annotated data. Most of the sentences are not relevant to a given privacy concern and a random sample will not contain enough positive examples to train a sensitive classifier. The combination of keywords and exemplar sentences helped improve the classification sensitivity to the small number of relevant sentences. State-of-the-art methods, such as those based on Deep Learning, offer high accuracy with additional complexity and the need for large amounts of annotated data. We instead decided to pursue a simple approach that can still produce useful results.

4.6. Evaluating the Sentence Classifiers

We decided to evaluate the performance of the classifiers described in Sections 4.1 through 4.5 (Fig. 1, lower right) in terms of sensitivity and specificity instead of precision and recall as is typically done for information retrieval. This is a consequence of us being more interested in minimizing false negatives than minimizing false positives. With privacy policies, the former represent obligations missed by the search which cannot easily be found otherwise while the latter can be eliminated via manual inspection. False negatives can represent legal liabilities especially for a new product or service.

Sensitivity and recall are both synonyms for the true positive rate which is the fraction of correctly classified positive items [24, 25]. Specificity is defined as the true negative rate (the fraction of correctly classified negative items) [24, 25]. Classifier evaluation in terms of precision is sensitive to class imbalances because changes in the test data class distribution will change the classifier's apparent performance [24, 26]. By only considering a single class of a binary classification, specificity and sensitivity do not suffer from this "class skew" [24, 26]. This separate focus on the performance for each class also allows for the consideration of "cost skew", the different cost associated with errors for each class [25].

As mentioned above, we are more concerned with false negatives than false positives; for privacy policy analysis, we believe that the cost associated with misclassifying sentences as not relevant is larger than that of misclassifying them as relevant. Using sensitivity and specificity will allow us to separate these two situations.

The similarity-based classifiers were evaluated using sentences drawn from the test data. For each classifier, sentences that contained the keywords were used to create a test set. The sentences selected and rejected by each classifier were manually examined to identify false positives and false negatives; we randomly sampled classifier results when the output was too large for human analysis.

We calculated the sensitivity and specificity for each classifier and used k-means clustering ($k = 3$) to partition these values (sensitivity clusters: 0, 1, and 2 and specificity clusters: 0, 1, and 2). We chose $k = 3$ to roughly separate the sensitivity and specificity values into three groups denoting "low," "medium," and "high" to facilitate their analysis. Further investigation is needed to determine whether this grouping of values is optimal but we believe that it is a good starting point. These clusters were then used to group the results and divide them into five cases:

(i) **Higher sensitivity and specificity** (sensitivity cluster 2 and specificity cluster 2).
(ii) **Higher specificity** (cluster 2 or "high") **with lower sensitivity** (cluster 0 "low" and cluster 1 or "medium").
(iii) **Higher sensitivity** (cluster 2 or "high") **with lower specificity** (cluster 0 or "low" and cluster 1 or "medium").
(iv) **Mixed lower sensitivity and specificity values** (all clusters except sensitivity cluster 2 and specificity clusters 2)
(v) **Undefined sensitivity or specificity**

5. Evaluation Results

We will now present the results of evaluating the method described Sections 4.1 through 4.5 using the process described in Section 4.6 for the use case of searching the RE 2013 privacy policy corpus for privacy concerns N4 through N20 described in the Notice section of the Marotta-Wurgler study. Concerns N1, N2, and N3 were deemed as being addressed outside of the posted privacy policies. Concerns N12 and N13 were combined because they represented concerns that were too similar to be resolved individually. Concern N21 was omitted because initial searches of the corpus with command-line tools did not find any candidate sentences for this privacy concern.

We began by dividing corpus into training and test data using a 66 %/33 % split as described in Section 4.1. This resulted in training data consisting of 1373 policy documents and test data with 687 policy documents. We then used spaCy to segment the 140 000 training sentences and 68 400 test sentences.

Exemplar sentences were drawn from the training data; we identified 56 sets of keywords and exemplar sentences using the process described in Section 4.3. We then used gensim to create a 400-concept VSM of the training data sentences using LSI (Section 4.4).

For the BERT-based approach, we used the large embedding provided by the Flair NLP framework [16, 27] which embeds text into a 1024-dimension space.

The keywords, exemplar sentences, and the vector space models were then used to create 112 classifiers using the Python-based gensim package [28] as described in Section 4.5, one for each combination of embedding approach, set of keywords, and associated exemplar sentences. An input sentence is selected if contains the keywords and if its cosine similarity with an exemplar sentence exceeds the preset selection threshold.

The classifiers were evaluated using the test data (Section 4.6). Of the 68 400 sentences in the test data, 5695 matched the keywords used by our classifiers and were used for evaluation. We calculated classifier sensitivity and specificity; their distributions are shown as histograms in Fig. 2. We used k-means clustering to partition the sensitivities and specificities into three clusters: "low," "medium," and "high"; their centroids and boundaries appear as vertical lines in Fig. 2. These clusters

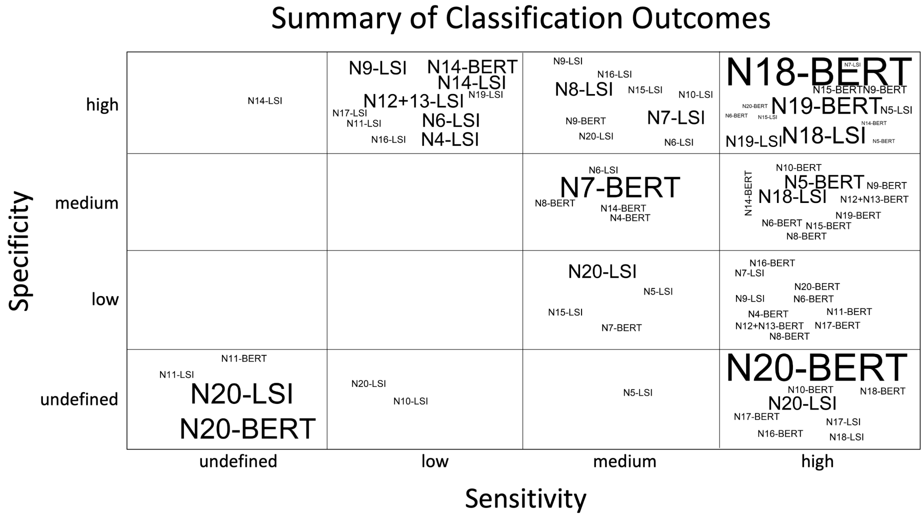


Figure 3: Results partitioned into tag clouds by their sensitivity and specificities. Each cell represents a separate tag cloud. The upper right cell contains results with the highest sensitivities and specificities, whereas the lower left cell has undefined sensitivities and specificities. Tag height is proportional to the number of associated classifiers. Tag position is not significant.

were used to group the results in Fig. 3 which were then divided into the five cases described above in Section 4.6 and in the rows of Table 2.

Case 1: Combined Higher Specificity and Sensitivity In addition to the results shown in Table 2, Table 3 gives details for the Case 1 classifiers which belonged to both sensitivity cluster 2 ("high sensitivity") and specificity cluster 2 ("high specificity"). This case contains 9 of the 17 concerns for which we applied our method and consisted of mostly relevant sentences with few missing.

Case 2: Higher Specificity with Lower Sensitivity These results consisted of mostly relevant sentences but the lower sensitivities meant some were missing. When combined with Case 1 above, 97 % of the test sentences were classified with high specificity; sentences identified as relevant to a privacy concern were likely to be relevant. These high specificity results corresponded to all of the concerns (N4 through N20) we considered.

Case 3: Lower Specificity with Higher Sensitivity Here, most of the relevant sentences were selected but many non-relevant ones were included as well. When combined with Case 1, 93 % of the test sentences contain both cases' keywords and be found with high sensitivity. These high sensitivity classifications also corresponded to all of the concerns (N4 through N20) we considered.

Cases 4 and 5 The last two rows of Table 2 show the results for *Case 4 (Lower Specificity and Lower Sensitivity)* and *Case 5 (Undefined Sensitivity or Specificity)*. The former were a mixture of relevant and non-relevant sentences. The latter's had undefined specificities or sensitivities which occurred when the test set lacked either positive or negative examples.

We can now address our research questions:

RQ1: Which of the Marotta-Wurgler notice-related privacy concerns can be identified using our simple sentence similarity-based approach?

Answer: Based on the measures that we've chosen for **RQ1** in Section 4, we are able to identify sentences from 9 of the 20 one notice-related privacy concerns given in the Marotta-Wurgler study. These privacy concerns are marked with a dagger in Table 1.

RQ2: How well does our similarity-based approach to identify privacy concerns in a privacy policy corpus?

Answer: The 9 notice-related concerns from Case 1 above (Section 5) were classified with an average sensitivity of 93.8 % and an average specificity of 98.3 %. When combined with the results of Case 2 whose sentences were identified with an average specificity of 97.8 % and an average sensitivity of 35.6 %, then sentences from 17 of the 21 notice-related privacy concerns can be identified with specificities of at least 81 %.

The implications of these results for large scale analyses of privacy policies are discussed in Section 6.

6. Discussion

Our goal is a means to analyze large privacy policy corpora with reduced manual effort by leveraging the semi-

Table 2: Privacy Policy Sentence Classification Results

Case	Specificity Level	Sensitivity Level	Candidate Sentences	Candidate % of Total	Found with LSI	Found with BERT	Specificity, % Min./Avg./Max.	Sensitivity, % Min./Avg./Max.
1	Higher	Higher	777	13.6	188	209	83.3 / 98.3 / 100	81.3 / 93.8 / 100
2	Higher	Lower	5458	95.8	1253	555	82.2 / 97.8 / 100	0.00 / 35.6 / 71.4
3	Lower	Higher	5000	87.8	113	4379	0.00 / 41.6 / 80.0	87.3 / 92.6 / 100
4	Lower	Lower	837	14.7	109	427	0.00 / 44.7 / 80.0	50.0 / 68.2 / 80.0
5	Undefined †		263	4.6	162	129	NA	NA

† Contains results with either unspecified sensitivities or specificities.

automated analysis of a small random sample that to develop classifiers that can be used to analyze the remainder. These classifiers are grounded in the corpus; we don't expect them to work with other corpora. Rather, we see the overall methodology as being the key contribution to the analysis large privacy policy corpora.

We used privacy concerns identified by Marotta-Wurgler to explore if we can combine keyword searches with manual selection of exemplar sentences to develop classifiers that find test set sentences similar to the exemplar sentences from the training set. If successful, this could allow for tagging of documents in a large corpus with a manual effort similar to that needed for a sample.

Our classification results fell into the five cases given in Section 5 that represent the different conditions under which sentences for the privacy concerns are identified. We will now discuss the results for these cases.

Case 1: Combined Higher Sensitivity and Specificity Table 3 reveals that for some of the privacy concerns, the use of keywords and exemplar sentences combined with the similarity-based classifiers allowed for test set sentence identification with few false negatives and false positives. This case represents an ideal situation that offers automation with little manual effort to extend Marotta-Wurgler style analyses to large privacy policy corpora.

Case 2: Higher Specificity with Lower Sensitivity This case mostly contains LSI-based results (Fig. 2, middle of top row). Here we identified relevant sentences, but not all of them. The average specificity of 97.8 % and the large fraction of test sentences with keywords suggest that the best improvements will come by increasing classifier sensitivity. This can come by increasing the default number (20) of exemplars suggested by LexRank. For example, there were 1348 test sentences for (N9) "collect + personal + information"; more exemplar sentences could have increased the matches and the sensitivity.

When combined with Case 1, 97 % of the Case 2 test sentences contain both cases' keywords and can be found with high specificity; sentences identified as relevant are likely to be relevant. These combined cases also contain all the concerns (N4 through N20) that we considered.

A more sophisticated classifier could improve sentence identification by training on a large manually curated data set. Alternatively, an active learning that prioritizes sentences for manual inspection, such as the one described by Yu and Menzies may be useful [19].

Case 3: Lower Specificity with Higher Sensitivity This case's average sensitivity of 92.6 % suggests that our approach may be useful for small numbers of sentences amenable to manual removal of false positives, especially for the higher specificity values (80 %).

A means of improving the specificity when there are many sentences, is to train better classifiers using a manually annotated data drawn from initial results. These new classifiers, could then be used with the similarity-based one to identify sentences with a higher specificity.

These results are dominated by the BERT-based classifiers implying that BERT-based features may be more advantageous than LSI-based ones by more consistently allowing for the use of secondary classifiers when the initial classification produces higher sensitivity results.

Case 4: Lower Specificity and Lower Sensitivity These results may be enough for the automated comparative analyses described above. For example, N6 ("pages + viewed") with LSI classifiers, and N7 ("collect+financial+information") with BERT classifiers had specificities of 77.9 % and 80.0 % and sensitivities of 79.8 % and 73.1 %, respectively. As in Case 2, better classifiers and active learning may improve the results.

Case 5: Undefined Sensitivity or Specificity Undefined sensitivities meant that there were no relevant test data items. For example, though the N14 keywords ("use + photograph + advertising − agree") matched three test sentences, all were correctly identified as not relevant. Undefined specificities meant that there were only relevant test set items; all sentences matching the keywords should be selected and our classifiers provided no benefit.

Application to Privacy Policy Analyses As discussed above, we rely on relevant keywords for each privacy concern; not all keywords lead to high sensitivity results. If a concern is best captured by multiple sets of keywords, each leading to classifications with differing sensitivities, then we will likely miss relevant sentences. However, we obtained high-specificities for the majority of the notice-related privacy concerns. While we cannot guarantee finding all relevant sentences, we can achieve situations

Table 3: Sentence Classification Results for Case 1

concern	keywords	VSM	exemplars	test set	sens	spec
N5	collect+browser +type	LSI	20	65	88.9	100
N5	collect +operating +system	LSI	16	65	90.4	100
N5	collect+operating +system	BERT	16	65	86.8	83.3
N6	your+browsing +history	BERT	20	9	87.5	100
N7	include+credit +history	LSI	20	16	83.3	100
N9	collect+medical +information	BERT	16	12	83.3	83.3
N9	collect+sensitive +information	BERT	20	16	92.9	100
N14	use+information +advertising	BERT	20	428	94.6	88.5
N15	advertisements +third+usage	BERT	11	12	90.9	100
N15	advertisements+third +visit+collect-agree	LSI	15	11	100	100
N15	advertisements+third +visit+collect-agree	BERT	15	11	90.0	100
N18	material+change+prior	LSI	20	22	89.5	100
N18	material+change+prior	BERT	20	22	100	100
N18	notice+change+notify	BERT	20	58	100	100
N18	notice+change+prior	BERT	20	68	100	100
N18	policy+change+notify	LSI	20	58	82.1	100
N18	policy+change+notify	BERT	20	58	96.4	100
N18	policy+change+prior	LSI	20	17	100	100
N19	material+change +accept	LSI	8	3	100	100
N19	material+change +accept	BERT	8	3	100	100
N19	notice+change +accept	LSI	16	20	81.3	100
N19	notice+change +accept	BERT	16	20	100	100
N19	policy+change +accept	LSI	20	28	100	100
N19	you+opt-in+change	BERT	1	2	100	100
N20	material+change +previous	BERT	3	6	100	100

where most of those identified are and their relative frequencies and distributions across a corpus can form the basis of large-scale analyses that yield valuable insights, such as differences across market segments, in a manner similar to Marotta-Wurgler's analysis.

We envisage iterative analyses, starting with keyword selection using text searches to verify that they lead to relevant sentences. Subsequent classification using exemplar sentences may uncover anomalies such as the absence of a privacy concern in a segment of the remaining corpus. Further investigation could lead to new keywords and exemplar sentences. Because its substantial qualitative nature, the concept of saturation should be used to end the iterations; it is reasonable to stop when no new information is found despite systematic attempts [29].

7. Limitations

The methodology we present is a starting point for further investigation of large scale semi-automated policy analyses. Despite our encouraging results, we have not solved the problem of semi-automated privacy policy analysis.

Our use of keywords is intended to mitigate the class imbalances that occur in a random samples of privacy policy sentences. We do not offer guidance on keyword selection or evaluation. We believe we have made reasonable choices. Further research could better determine their selection or eliminate them altogether.

We rely on similarity to exemplar sentences to reduce the need for human-annotated ground truth and to keep our baseline method simple while producing useful results. We have necessarily forsaken the performance of state-of-the-art methods. Additional studies could explore the tradeoffs of using more complex methods.

We assumed sentences are at the appropriate level of granularity for privacy policy analyses. This may not hold for all privacy concerns, but this was both convenient for our analysis and worked reasonably well for those described by Marotta-Wurgler's study.

For simplicity, we used a threshold for the cosine similarity of 0.5 with the intuition that above this value two vectors are more mutually aligned than not. In practice, classifier receiver operating characteristic (ROC) curves could be used to select the thresholds.

The clustering of sensitivities and specificities into "low," "medium," and "high" (Fig. 3) and their grouping into "lower" and "higher" (see Table 2 and Section 5) result from our choice of clustering algorithm and our need to partition our results for analysis and discussion. This is a qualitative interpretation similar to inter-rater comparisons [30]. Further work is necessary to determine optimal result clustering and labels in practice.

8. Summary and Future Work

We presented a semi-automated methodology for extending Marotta-Wurgler style analyses to large privacy policy corpora and provided an initial feasibility evaluation with an existing privacy policy corpus. Our methodology begins with a keyword-based search followed by extractive summarization to find the most representative sentences in a policy corpus. They are then manually inspected to find exemplars for similarity-based classifiers that identify other relevant sentences. The classifiers identified sentences addressing concerns with varying results. For those identifiable with higher sensitivity and specificity, the manual effort of analyzing a small subset of a large privacy policy corpus could lead to automated analysis of the rest. This condition accounted for 9 of the 17 privacy concerns for which we used our methodology.

When sentences were identified with higher sensitivity but lower specificity, the results were dominated by the BERT-based similarity classifiers. The possibility of improving the specificity suggests an advantage of BERT over LSI. Here the non-relevant sentences could be removed manually or by using another classifier trained on a subset of the corpus. When combined with the higher sensitivity and specificity results, we were able to identify sentences from all seventeen privacy concerns with which we evaluated our methodology.

Our future efforts will focus on increasing the number of privacy concerns identified while maintaining the aesthetic of simplicity and minimal training data requirements. We will investigate using our method to create training sets for additional downstream classifiers as well as active learning approaches. We will also evaluate its use on a much larger privacy policy corpus.

References

[1] S. Cowley, "Equifax to pay at least $650 million in Largest-Ever data breach settlement," *The New York Times*, July 2019.

[2] E. C. Baig, "Facebook fined $5 billion by FTC, must update and adopt new privacy, security measures," *USA Today*, July 2019.

[3] A. K. Massey, J. Eisenstein, A. I. Anton, and P. P. Swire, "Automated text mining for requirements analysis of policy documents," *2013 21st IEEE Intl Requirements Eng. Conf., RE 2013 - Proc.*, pp. 4–13, 2013.

[4] F. Marotta-Wurgler, "Understanding privacy policies: Content, Self-Regulation, and Markets," Tech. Rep. 16-18, New York University School of Law, 2016.

[5] L. Netcraft, "January 2019 web server survey | netcraft." `https://news.netcraft.com/archives/2019/01/24/january-2019-web-server-survey.html`, 2019. Accessed: 2019-8-19.

[6] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *1*, vol. 22, pp. 457–479, Dec. 2004.

[7] R. Feldman, "Text mining," in *Handbook of Data Mining and Knowledge Discovery* (W. Klosgen and J. M. Zytkow, eds.), pp. 749–757, Oxford University Press, 2002.

[8] J. Lin and D. Gunopulos, "Dimensionality reduction by random projection and latent semantic indexing," *Proc. Text Mining Workshop, at the 3rd SIAM Intl. Conf. Data Mining*, 2003.

[9] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press, 1999.

[10] B. Liu, *Web Data Mining*. Berlin: Springer-Verlag, second edi ed., 2011.

[11] S. Deerwester, S. T. Dumais, and T. K. Landauer, "Indexing by latent semantic analysis," *J American Soc. for Inform. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[12] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse Process.*, vol. 25, pp. 259–284, Jan. 1998.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Adv. Neural Inform. Process. Sys. 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 3111–3119, Curran Associates, Inc., 2013.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[15] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," Feb. 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Oct. 2018.

[17] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, *Semantic Similarity from Natural Language and Ontology Analysis*, vol. 8 of *Synthesis Lectures Human Language Technologies*. Morgan & Claypool Publishers, 2015.

[18] M. R. Grossman, G. V. Cormack, and A. Roegiest, "TREC 2016 total recall track overview," in *TREC*, 2016.

[19] Z. Yu and T. Menzies, "Total recall, language processing, and software engineering," Aug. 2018.

[20] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th USENIX Security Symposium*, pp. 531–548, 2018.

[21] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, "We value your privacy ... now take some cookies: Measuring the GDPR's impact on web privacy," in *Proceedings 2019 Network and Distributed System Security Symposium*, (Reston, VA), Internet Society, 2019.

[22] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the GDPR," *Proc. on Privacy Enhancing Tech.*, vol. 2020, pp. 47–64, Jan. 2020.

[23] R. B. Bradford, "An empirical study of required dimensionality for large-scale latent semantic indexing applications," in *17th ACM conference on Information and knowledge mining*, (Napa Valley, California), pp. 153–162, ACM Press, 2008.

[24] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," Tech. Rep. HPL-2003-4, HP Laboratories, Jan. 2003.

[25] J. Korst, V. Pronk, M. Barbieri, and S. Consoli, "Introduction to classification algorithms and their performance analysis using medical examples," in *Data Science for Healthcare: Methodologies and Applications* (S. Consoli, D. Reforgiato Recupero, and M. Petković, eds.), pp. 39–73, Cham: Springer International Publishing, 2019.

[26] F. M. Harper, D. Moy, and J. A. Konstan, "Facts or friends?: distinguishing informational and conversational questions in social Q&A sites," in *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Assoc. for Computing Machinery, 2009.

[27] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An Easy-to-Use framework for State-of-the-Art NLP," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[28] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta), pp. 45–50, ELRA, May 2010.

[29] S. B. Merriam and E. J. Tisdell, *Qualitative Research*. John Wiley & Sons, Inc., 4th editio ed., 2016.

[30] K. A. Hallgren, "Computing Inter-Rater reliability for observational data: An overview and tutorial," *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.

17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020
Paper N° C002410
Registration Code: S-A01913

COLLAPSE ESTIMATES OF U.S. CODE-COMPLIANT STEEL FRAMES AND IMPLICATIONS FOR AN ASCE 41 ASSESSMENT

M.S. Speicher[(1)], K.K.F. Wong[(2)], J. Dukes[(3)]

[(1)] *Research Structural Engineer, National Institute of Standards and Technology, USA, matthew.speicher@nist.gov*
[(2)] *Research Structural Engineer, National Institute of Standards and Technology, USA, kevin.wong@nist.gov*
[(3)] *Research Structural Engineer, National Institute of Standards and Technology, USA, jazalyn.dukes@nist.gov*
…

Abstract

ASCE 41 is a standard that contains performance-based engineering procedures sometimes used to assess and retrofit existing structures. In 2015, the U.S. National Institute of Standards and Technology completed a study investigating the relationship between ASCE 41 and traditional new building design standards. A key observation from this study was that there are inconsistencies between the two approaches, some of which may be caused by suspected conservatism in the ASCE 41. To further investigate this relationship, this study will present the results of a FEMA P695 assessment of a set of six steel special moment frames. The goal is to verify that the ASCE 7 design intent of no greater than 10 % probability of collapse given a risk-targeted maximum considered earthquake is being met. The effects of various modeling assumptions, such as backbone curves, damping, and p-delta columns, are discussed. The sensitivity of these modeling approaches is summarized, and it is found that consideration of the beam-slab composite action is most significant. Furthermore, the tendency for conservative assumptions for one component resulting in non-conservative results for another component is highlighted. It is expected this study will provide useful and timely information to engineers and standards committee members charged with improving performance-based seismic design approaches.

Keywords: Performance-based design, ASCE 41, collapse, FEMA P695, steel moment frames

Speicher, Matthew; Wong, Kevin K F; Dukes, Jazalyn. "Collapse estimates of U.S. code-compliant steel frames and implications for an ASCE 41 assessment." Presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

One general goal of performance-based seismic design is to allow an engineer to make design decisions outside the typically prescriptive requirements of traditional building codes and standards. In theory, this gives the engineer freedom to make creative and/or efficient choices that produce a well-performing and cost-effective structure. Currently in the U.S., the American Society of Civil Engineers (ASCE) standard 41 titled *Seismic Evaluation and Retrofit of Existing Buildings* is the consensus document used for both assessment of existing buildings and alternative, performance-based design of new buildings [1]. This standard is often simply referred to as ASCE 41. In contrast, most new buildings are constructed using design procedures from ASCE 7, *Minimum Design Loads and Associated Criteria for Buildings and other Structures* [2] and associated materials specific specifications such as the American Institute of Steel Construction (AISC) publication 360 titled *Specification for Structural Steel Buildings* [3] and publication 341 titled *Seismic Provisions for Structural Steel Buildings* [4].

Given these different available approaches for the design of new buildings, researchers at the National Institute of Standards and Technology (NIST) conducted a study investigating the relationship between ASCE 41 and the corresponding new building standards/specifications. The findings of the NIST investigation were published in a comprehensive series of reports [5-7] and journal papers [8-11] over the last several years. An over-arching conclusion from the NIST investigation is that ASCE 41 tends to be overly-conservative and new provisions should be considered to better align ASCE 41 with new building standards/specification. However, the authors noted even though their study showed, in many cases, a building designed with the current conventional standards and specifications would fail an ASCE 41 assessment, there is reason to question whether this is, in fact, a misleading conclusion. In other words, though the buildings were properly designed to meet or exceed the requirements of ASCE 7, AISC 360, and AISC 341, there is no guarantee the building would actually meet the performance intent of the building code, which is a no greater than 10 % probability of collapse given a risk-targeted “maximum considered earthquake.”

Therefore, this paper investigates the NIST-designed special moment frames to more fully-contextualize the conclusions made in previous studies. This is done using a rigorous methodology established by the Federal Emergency Management Agency (FEMA) publication P695 titled *Quantification of Building Seismic Performance Factors* [12]. To quantify the performance, FEMA P695 provides an approach that incrementally increases the earthquake intensity to determine a structure’s collapse behavior. This process is repeated for a set of 44 earthquake records (22 pairs), and then the distribution of collapses is plotted relative to intensity measure to create a lognormal cumulative distribution function know as a fragility curve. The basis of this methodology is commonly referred to as incremental dynamic analysis (IDA). To facilitate efficient deployment of IDA, new nonlinear building models are created in OpenSees [13]. The OpenSees models also allow the use of current state-of-the-art nonlinear degradation models, which are important when considering collapse-level shaking.

2. Methodology

The suite of archetype steel buildings come from Harris and Speicher [5]. These buildings have a special moment frame (SMF) in the east-west direction and a special concentrically braced frame in the north-south direction. As mentioned previously, only the special moment frames are investigated in this study. For each building height, the frames were designed using the equivalent lateral force (ELF) procedure and the modal response spectrum analysis (RSA) procedure. Fig. 1 gives the general floor framing layout. Fig. 2 gives the member sizes for the moment frames, including the reduced beam section (RBS) dimensions. The floors are assumed to be composite slabs with 5 in (12.7 cm.) concrete slab. The stiffness of the slab is considered for both the design and assessment of the building, but composite-slab interaction is not considered when modeling the beam hinges for the baseline analysis. The effects of composite slab interaction are discussed in the follow-on NIST report related to this work [14]. The wide-flange sections are assumed to be A992 Grade 50 steel. Further design details can be found in Harris and Speicher [5].

Details of the nonlinear models used are shown in Fig. 3. The models used in Harris and Speicher [5] were created in Perform-3D [15]. Given the computational expense of running IDA, the three-dimensional Perform-3D models were converted into two-dimensional OpenSees models. This was advantageous both in terms of efficiency and enabling the use of nonlinear deterioration models that reflect state-of-the-art research. A rigorous comparison between the three-dimensional Perform-3D model and the two-dimensional OpenSees model and the results show reasonable agreement (see [14] for full details).

The moment frames have reduced beam section (RBS) connections with stiffness defined using a prismatic cross-section over the length of the RBS (length *b* in Fig. 3). The width of the RBS was approximated as the actual RBS width at *b*/3 away from the center of the RBS. A nonlinear rotational spring was placed at center of the RBS and was assigned a stiffness of 10 times that of the unreduced beam. Since this spring is in series with the elastic beam elements, the elastic beam stiffness was increased to give an overall beam (including the RBS and nonlinear spring) stiffness equal to that without the nonlinear spring (see [14] Appendix D for further discussion). The nonlinear spring was assigned the modified Ibarra Medina Krawinkler (IMK) model which uses the OpenSees Bilin material model. The force-deformation parameters followed recommendations from Lignos and Krawinkler [16].

For the columns, nonlinear springs were placed half the depth of the column (d / 2) away from the face of the beam. These nonlinear springs followed the same approach as the beams. The force-deformation parameters followed the recommendations from NIST [17] using a monotonic backbone plus accompanying degradation. At the intersection of each beam and column, the panel zones were modeled explicitly using an approach outlined by Krawinkler [18] – which includes a set of “rigid” elements with pinned connections tied together with nonlinear rotational spring in one corner. The spring parameters were based on fundamental mechanics. Though the column splice was designed at 4 ft (1.2 m) above the beam-to-column joint, this was ignored in the model.

To account for the gravity frame in the model, a P-delta column is added and connected to the moment frame with rigid links at each story. The moment of inertia of the P-delta column was taken as tributary sum of the moment of inertias of the gravity and SCBF columns to account for the stiffening effect of the frame outside of the SMF. For simplicity, the P-delta column was assumed elastic along the height of the building. Additionally, the nonlinear models are given 3 % modal damping for all modes and 0.3 % stiffness proportional damping to damp out spurious higher modes. This damping selection aligns the OpenSees models with the models used in Perform-3D by Harris and Speicher [5].

IDA was performed to determine each frame’s collapse fragility. The IDA followed guidelines set forth in FEMA P695. A suite of 44 ground motion records (22 pairs) were used in the IDA, which are referred to as the “far-field” set in FEMA P695. The intensity measure used for the IDA is the median spectral acceleration of the earthquake suite at the fundamental period of the building, C_uT_a. Collapse is defined as any story drift exceeding 7.5 % or when the analysis failed to converge.

To determine the fragility curve from the IDA results, the probability of collapse at intensity measure x, P(C | IM = x), can be calculated in Eq. (1) as follows:

$$P\left(C \mid IM = x\right) = \Phi\left(\frac{\ln\left(x/\theta\right)}{\beta}\right) \tag{1}$$

where the normal cumulative distribution function (CDF) is denoted by Φ, the median of the fragility function is denoted by θ, and the standard deviation of ln(IM) is denoted by β. Eqs. (2) and (3) show the formulation of the fragility function estimators (designated with the “hat” marking) over *n* number of earthquakes, which are method of moments estimators of a normal distribution [19].

$$\ln\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\ln\left(IM_i\right) \tag{2}$$

$$\hat{\beta} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln\left(IM_i/\hat{\theta}\right)\right)^2} \tag{3}$$

The analyses were performed in part using the parallel version of OpenSees, OpenSeesMP, on the Extreme Science and Engineering Discovery Environment (XSEDE) platform [20] and on local NIST computers.

After the IDA is completed, collapse margin ratios (CMRs) can be determined directly from the fragility curves. The intensity measure of interest, S_T, is defined as the median spectral acceleration of the record set at period *T*. The *CMR* is calculated as follows:

$$CMR = \frac{\hat{S}_{CT}}{S_{MT}} \tag{4}$$

where $\hat{S}_{CT}$ is the median collapse intensity (i.e., the value of S_T that results in 50 % of the ground motions causing collapse) and S_{MT} is the value of the MCE$_R$ spectrum at period *T*.

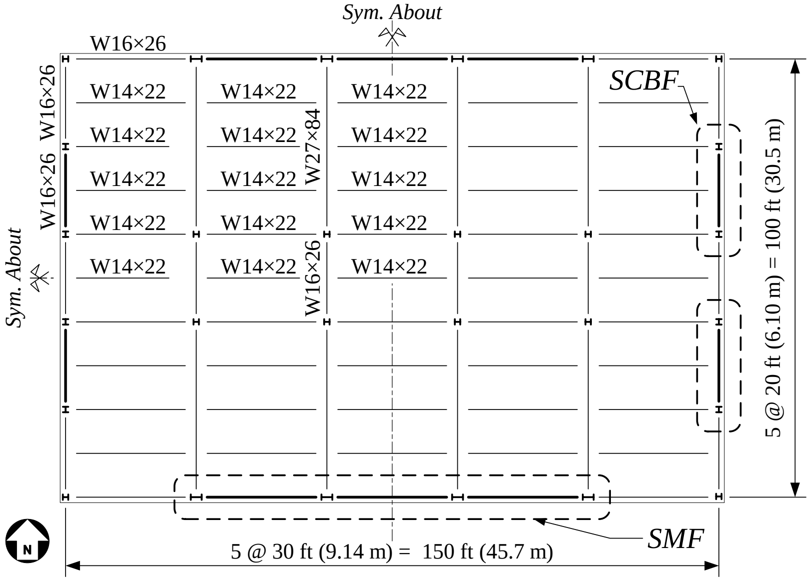
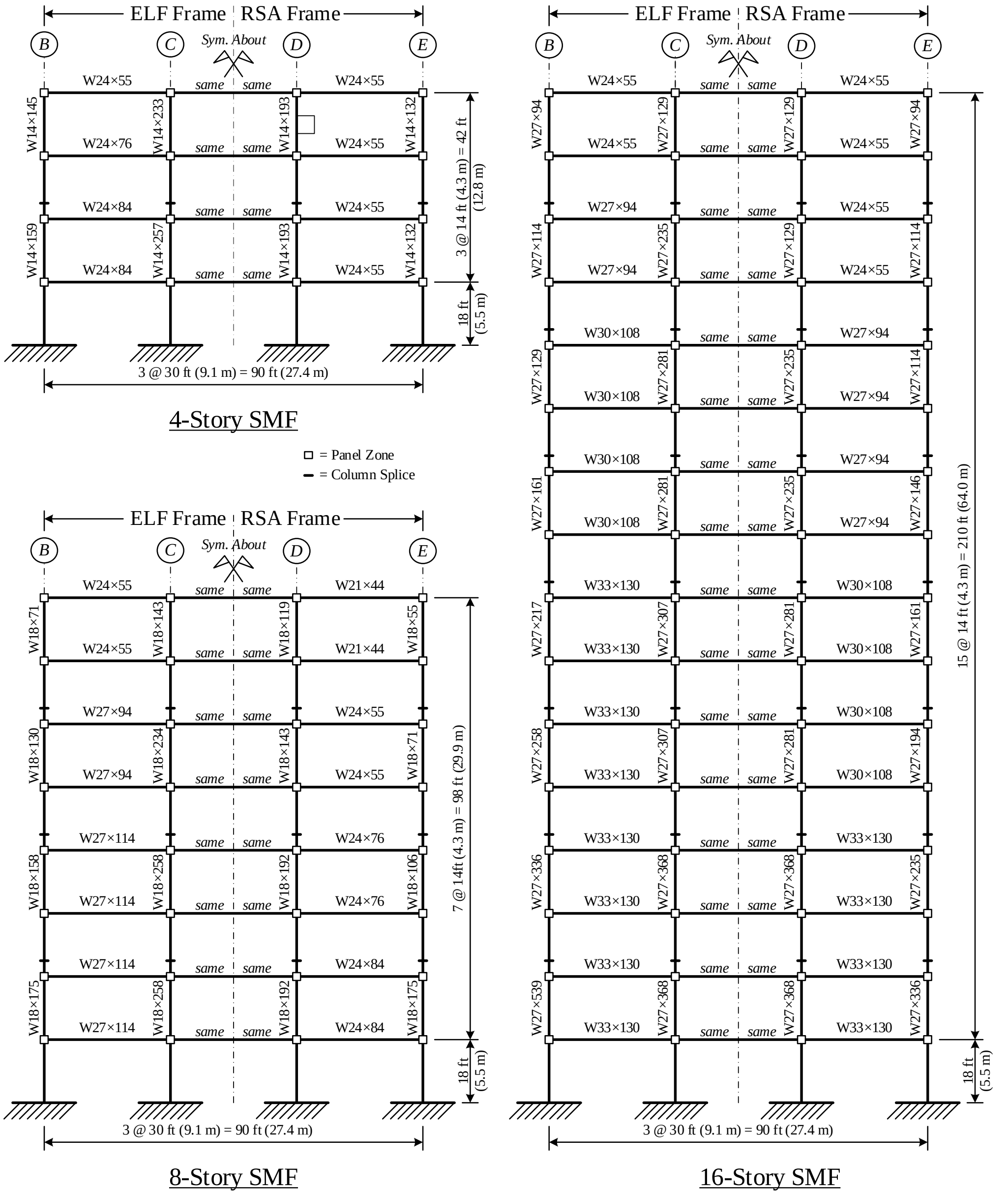


Fig. 1 – Typical building floor plan showing the structural framing layout



RBS Dimensions:

W24×55	*a* = 3.75", *b* = 16", *c* = 1.75"	W27×114	*a* = 5.25", *b* = 18", *c* = 2.50"
W24×76	*a* = 4.50", *b* = 16", *c* = 2.25"	W30×108	*a* = 5.25", *b* = 20", *c* = 2.50"
W24×84	*a* = 4.75", *b* = 16", *c* = 2.25"	W33×130	*a* = 5.75", *b* = 22", *c* = 2.75"
W27×94	*a* = 5.00", *b* = 18", *c* = 2.50"		

Note: steel section sizes provided in U.S. standard notation

Fig. 2 – The six SMF designs investigated in this study [14]

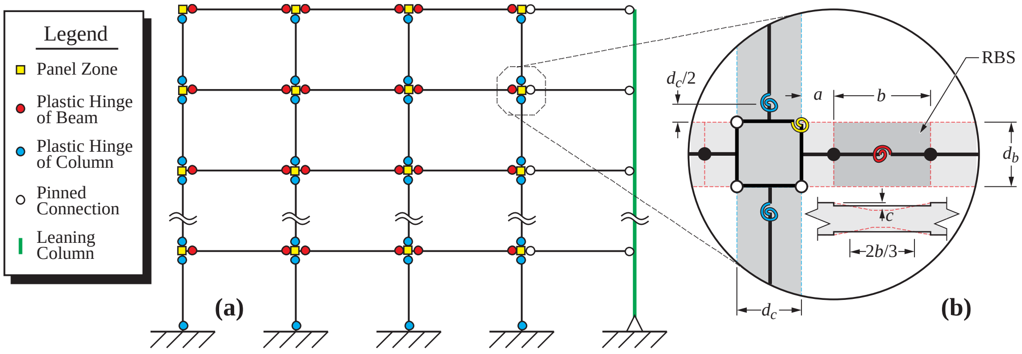


Fig. 3 – OpenSees modeling details for the special moment frames

3. Results

The results of the IDA are shown in terms of maximum story drift ratio versus normalized intensity measure S_T / S_{MT} and in terms of a fitted cumulative distribution function. Fig. 4 shows the IDA curves for the 4, 8, and 16-story ELF and RSA-designed frames. The IDA curves illustrate the progression of nonlinear response as the intensity of each earthquake is increased. Recall that the intensity measure, S_T, is the median spectral acceleration of the suite of earthquakes, not the spectral acceleration for any individual earthquake. Fig. 5 shows the fragility curves for the 4, 8, and 16-story ELF and RSA-designed frames. Since S_{MT} is the spectral acceleration at the MCE_R level, it can be quickly deduced what the probability of collapse is given an MCE_R. However, spectral shape and other variability has not yet been accounted for, which is necessary to complete the FEMA P695 process. Therefore, the median spectral acceleration (P(Collapse) = 0.5) is the value determined from the fragility curve; this value is referred to as the collapse margin ratio, *CMR*.

4. Discussion

To complete the FEMA P695 assessment, the *CMR* from Fig. 5 needs to be adjusted to account for the frequency content of the earthquake. The period-based ductility factor, μ_T, considers the period elongation caused by yielding in the structure. The μ_T, is obtained from pushover plots presented in [14]. The spectral shape factor, *SSF*, then uses the μ_T and accounts for the difference between the frequency content of rare ground motions versus less rare ground motions. Finally, the adjusted collapse margin ratio, *ACMR*, can be calculated as follows in Eq. (5):

$$ACMR = SSF \times CMR \quad (5)$$

The *ACMR* is then compared to an acceptable collapse margin ratio, given the desired collapse probability target and the total system uncertainty. The total system uncertainty values are largely based on judgement and explained in FEMA P695. The total collapse uncertainty, β_{TOT}, is calculated using the square-root-of-the-sum-of-the-squares (SRSS) of the (1) the record-to-record uncertainty, β_{RTR}, (2) the design requirements-related collapse uncertainty, β_{DR}, (3) the test data-related uncertainty, β_{TD}, and (4) the modeling-related uncertainty, β_{MDL}. For β_{RTR}, a fixed value of 0.40 is assumed because FEMA P695 states that this is appropriate for the performance evaluation of systems with significant period elongation, which is expected to be true for special steel moment frames designed with a response modification coefficient, *R*, of 8. For β_{DR}, it is assumed there is high confidence in the basis of the design requirements and medium confidence in terms in completeness and robustness. Therefore, the result is a "good" rating, and a β_{DR} is set to 0.2. For

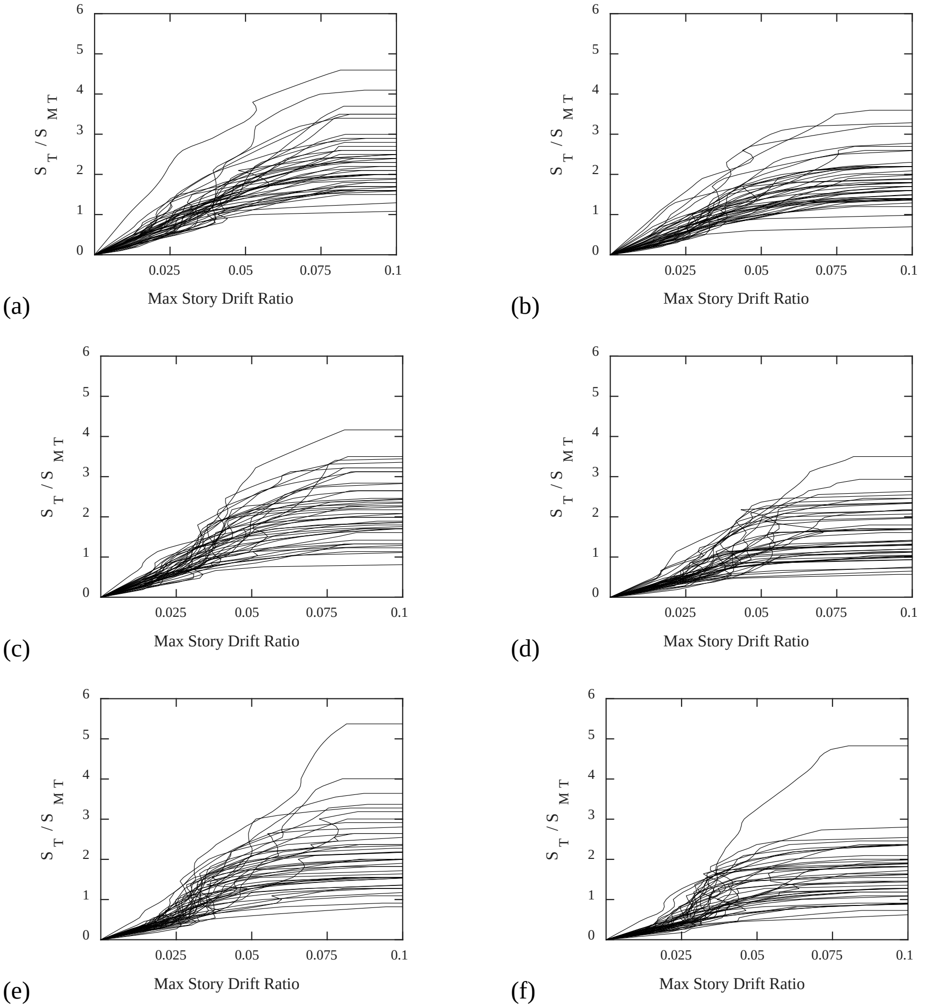


Fig. 4 – Incremental dynamic analysis curves for the (a) 4-story ELF, (b) 4-story RSA, (c) 8-story ELF, (d) 8-story RSA, (e) 16-story ELF, and (f) 16-story RSA-designed frames

β_{TD}, it is assumed there is high confidence the test results but a medium level of completeness and robustness of test results. This also gives a "good" rating, and a β_{TD} is set to 0.2. Lastly, for β_{MDL}, it is assumed that the accuracy and robustness of the models is high, but the representation of collapse characteristics is medium. This again gives a "good" rating, and a value of 0.2 is assigned. Now combining all the uncertainties using the SRSS, the total system collapse uncertainty is found to be 0.53. Using this total uncertainty number enables the determination of an acceptable value of *ACMR* from FEMA P695 Table 7-3, which is 1.96 for 10 % probability of collapse and 1.56 for 20 % probability of collapse. For a suite of new archetype building designs, the generally accepted collapse probability target is 10 % or less, given a $\mathrm{MCE_R}$. It is recognized that for an individual building design, probability of collapse, given a $\mathrm{MCE_R}$, may reach as high as 20 %, which is still considered acceptable. Therefore, since a limited suite of archetypes are investigated in this

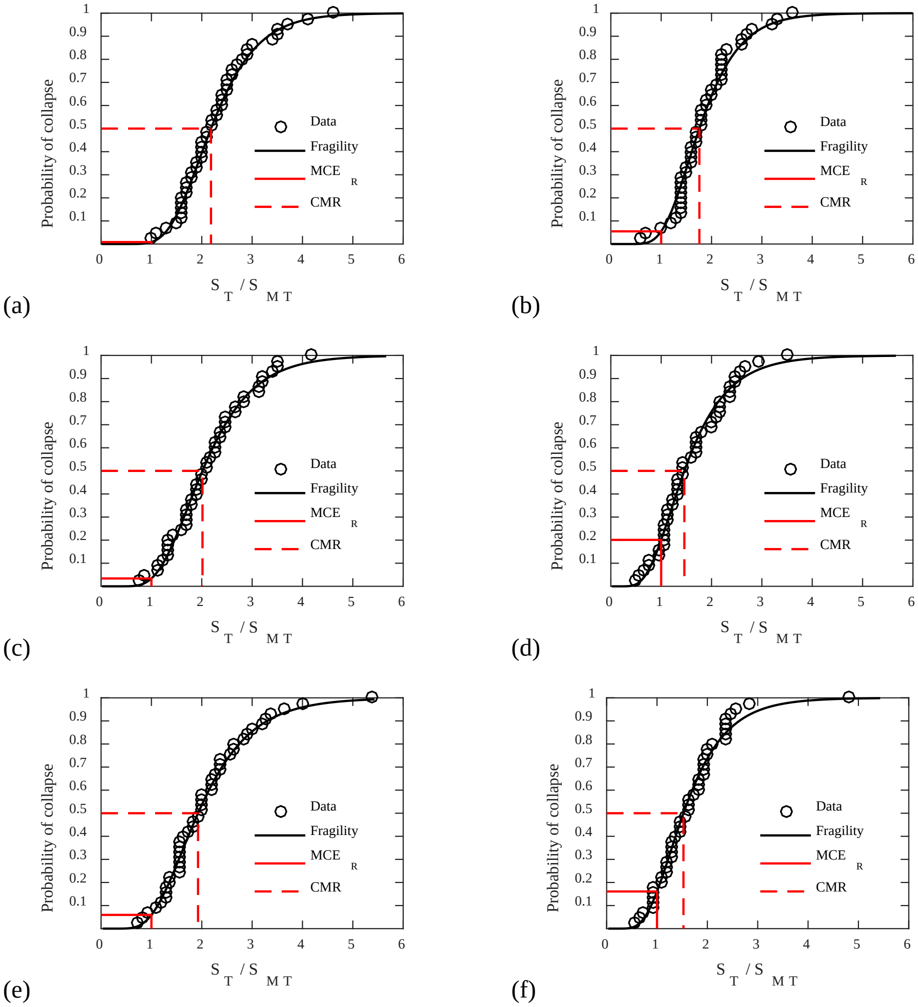


Fig. 5 – Fitted fragility curves for the (a) 4-story ELF, (b) 4-story RSA, (c) 8-story ELF, (d) 8-story RSA, (e) 16-story ELF, and (f) 16-story RSA-designed frames

study, the targets of $ACMR_{10\%}$ and $ACMR_{20\%}$ are both presented.

Table 1 gives a summary of the FEMA P695 assessment with all relevant values. Note that the FEMA P695 assessment considers the building performance as a whole, not as individual components as done in ASCE 41. Table 2 gives a summary of a nonlinear performance assessment using ASCE 41 reported in [8]. The nonlinear dynamic assessment procedure in ASCE 41 is the most comparable to a collapse assessment using the FEMA P695 methodology. Four out of the six buildings do not pass the ASCE 41 nonlinear dynamic assessment due to deficient beam-to-column connections. The columns and the panel zones also fail in some of the buildings. In contrast, all of the buildings pass the nonlinear FEMA P695 assessment at the more stringent 10 % probability of collapse threshold. This supports the conclusion made in Harris and Speicher [5], in which the ASCE 41 nonlinear assessment methodology is overly-conservative. As suggested in a follow-on set of journal papers by Speicher and Harris [9-11] and Maison and Speicher [21], ASCE 41 should investigate converting the acceptance criteria to a more robust measure that is not dependent on loading history. Acceptance criteria based on cumulative ductility demands (e.g., energy-based) could be especially sensible for well-performing systems that meet current building code provisions. Older, more archaic, systems may still need to be assessed with the current ASCE 41 approach, given that failure modes may not be as well-known or controlled. Regardless, further research should explore updating the acceptance criteria in performance-based design to ensure robust measurement science.

Table 1 – Summary of collapse performance evaluation for the suite special moment frames [14]

	Assessment parameters					Acceptance check (Pass/Fail = P/F)					
Building	Static Ω	CMR	μ_T	SSF	ACMR	$ACMR_{10\%}$	P/F ratio	P/F	$ACMR_{20\%}$	P/F ratio	P/F
04-ELF	2.78	2.18	5.4	1.48	3.22	1.96	1.64	**Pass**	1.56	2.06	**Pass**
04-RSA	2.22	1.76	4.5	1.43	2.51	1.96	1.28	**Pass**	1.56	1.61	**Pass**
08-ELF	2.61	2.01	4.1	1.40	2.83	1.96	1.44	**Pass**	1.56	1.81	**Pass**
08-RSA	1.85	1.46	3.1	1.35	1.97	1.96	1.01	**Pass**	1.56	1.27	**Pass**
16-ELF	1.92	1.93	3.6	1.38	2.65	1.96	1.35	**Pass**	1.56	1.70	**Pass**
16-RSA	1.68	1.53	3.7	1.38	2.11	1.96	1.08	**Pass**	1.56	1.35	**Pass**

Table 2 – Summary of predicted component performance by the nonlinear procedures for the collapse (CP) structural performance level (SPL) at the basic safety earthquake (BSE)-2 earthquake hazard level (EHL) for each archetype building (adapted from [8])

Building Height	Design Procedure	Nonlinear Static			Design Procedure	Nonlinear Dynamic (mean)		
		BC	CM	PZ		BC	CM	PZ
4-story	ELF	Pass	Pass	Pass	ELF	Pass	Pass	Pass
	RSA	**Fail**	Pass	Pass	RSA	**Fail**	Pass	Pass
8-story	ELF	Pass	**Fail**	Pass	ELF	**Fail**	**Fail**	Pass
	RSA	Pass	**Fail**	Pass	RSA	**Fail**	**Fail**	**Fail**
16-story	ELF	Pass	Pass	Pass	ELF	Pass	Pass	Pass
	RSA	Pass	Pass	Pass	RSA	**Fail**	**Fail**	Pass

Note: BC = beam-to-column connection, CM = column member, PZ = panel zone

5. Conclusions

A set of steel special moment frames designed with current building standards and then assessed using ASCE 41 were shown to be deficient in previous investigations. To determine whether the building designs are indeed deficient, this paper presents the results of a FEMA P695 investigation. The results demonstrate that the current code-complying designed frames have less than or equal to a 10 % probability of collapse, given a risk-targeted maximum considered earthquake, for all cases assessed. This suggests ASCE 41 is overly-conservative in its assessment criteria. It is recommended that further study be done to address the conservatism of ASCE 41 implied in these results. It is then argued that a potential remedy is to migrate ASCE 41 assessment criteria into a cumulative-based or energy-based measure. This may be especially relevant for current code-conforming buildings or when using ASCE 41 as a performance-based alternative for new building design.

6. Disclaimer and Acknowledgements

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. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) through allocation MSS170023 supported by NSF grant number ACI-1548562.

The policy of the National Institute of Standards and Technology 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. This paper is an official contribution of the U.S. National Institute of Standards and Technology; not subject to copyright in the United States.

7. References

[1] ASCE (2017): *Seismic Evaluation and Retrofit of Existing Buildings*, *ASCE/SEI 41-17*, American Society of Civil Engineers: Reston, VA.

[2] ASCE (2017): *Minimum Design Loads for Buildings and Other Structures*, *ASCE/SEI 7-16*, American Society of Civil Engineers: Reston, VA.

[3] AISC (2016): *Specification for Structural Steel Buildings*, *ANSI/AISC 360-16*, American Institute of Steel Construction: Chicago, IL.

[4] AISC (2016): *Seismic Provisions for Structural Steel Buildings*, *ANSI/AISC 341-16*, American Institute of Steel Construction: Chicago, IL.

[5] Harris 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. https://doi.org/10.6028/NIST.TN.1863-1

[6] Harris JL, Speicher MS (2015): *Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings Volume 2: Special Concentrically Braced Frames*, NIST Technical Note 1863-2, National Institute of Standards and Technology: Gaithersburg, MD. https://doi.org/10.6028/NIST.TN.1863-2

[7] Harris JL, Speicher MS (2015): *Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings Volume 3: Eccentrically Braced Frames*, NIST Technical Note 1868-3, National Institute of Standards and Technology: Gaithersburg, MD. https://doi.org/10.6028/NIST.TN.1863-3

[8] Harris J, Speicher MS (2018): Assessment of Performance-Based Seismic Design Methods in ASCE 41 for New Steel Buildings: Special Moment Frames. *Earthquake Spectra*. 34(3): p. 977-999. https://doi.org/10.1193/050117EQS079EP.

[9] Speicher MS, Harris JL (2016): Collapse Prevention Seismic Performance Assessment of New Eccentrically Braced Frames using ASCE 41. *Engineering Structures*. 117: p. 344-357. https://doi.org/10.1016/j.engstruct.2016.02.018.

[10] Speicher MS, Harris JL (2016): Collapse prevention seismic performance assessment of new special concentrically braced frames using ASCE 41. *Engineering Structures*. 126: p. 652-666. https://doi.org/10.1016/j.engstruct.2016.07.064.

[11] Speicher MS, Harris JL (2018): Collapse Prevention seismic performance assessment of new buckling-restrained braced frames using ASCE 41. *Engineering Structures*. 164: p. 274-289. https://doi.org/10.1016/j.engstruct.2018.01.067.

[12] FEMA (2009): *Quantification of Building Seismic Performance Factors*, *FEMA P695*, Federal Emergency Managament Agency: Washington, D.C.

[13] McKenna F, Fenves GL (2016): *OpenSees command language manual. Version 2.5.0*, Pacific Earthquake Engineering Research Center: Berkeley, CA.

[14] Speicher MS, Dukes JD, Wong KKF (2020): *Collapse Risk of Steel Special Moment Frames per FEMA P695*, 2084, National Institute of Standards and Technology: Gaithersburg, MD. https://doi.org/10.6028/NIST.TN.2084

[15] CSI (2013): *Nonlinear Analysis and Performance Assessment for 3D Structures, PERFORM 3D*, Computers and Structures, Inc.: Berkeley, CA.

[16] Lignos DG, Krawinkler H (2011): Deterioration modeling of steel components in support of collapse prediction of steel moment frames under earthquake loading. *Journal of Structural Engineering*. 137(11): p. 1291-1302. https://doi.org/10.1061/(ASCE)ST.1943-541X.0000376.

[17] NIST (2017): *Guidelines for Nonlinear Structural Analysis for Design of Buildings, Part IIa – Steel Moment Frames, prepared by the Applied Technology Council for the National Institute of Technology and Standards*, NIST GCR 17-917-46v2: Gaithersburg, MD. https://doi.org/10.6028/NIST.GCR.17-917-46v2

[18] Krawinkler H (1978): Shear in Beam-Column Joints in Seismic Design of Steel Frames. *Engineering Journal*. 15(3): p. 82-91.

[19] Baker JW (2015): Efficient Analytical Fragility Function Fitting Using Dynamic Structural Analysis. *Earthquake Spectra*. 31(1): p. 579-599. https://doi.org/10.1193/021113eqs025m.

[20] Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, Roskies R, Scott JR, Wilkins-Diehr N (2014): XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering*. 16(5): p. 62-74. https://doi.org/10.1109/MCSE.2014.80.

[21] Maison BF, Speicher MS (2016): Loading Protocols for ASCE 41 Backbone Curves. *Earthquake Spectra*. 32(4): p. 1-20. https://doi.org/10.1193/010816EQS007EP.

CALIBRATION REPRODUCIBILITY OF MICROFORM GEOMETRY FOR NIST STANDARD ROCKWELL HARDNESS DIAMOND INDENTERS

J. Song[1], T.B. Renegar[2], S.R. Low[3], C.R. Beauchamp[4]

[1] Physical Measurement Laboratory, NIST, USA, song@nist.gov
[2] Physical Measurement Laboratory, NIST, USA, thomas.renegar@nist.gov
[3] Material Measurement Laboratory, NIST, USA, samuel.low@nist.gov
[4] Material Measurement Laboratory, NIST, USA, carlos.beauchamp@nist.gov

Abstract:

Stability and reproducibility of national and international Rockwell hardness scales largely depend on the calibration reproducibility of the microform geometry of the standard Rockwell hardness diamond indenters. From 1995 to 2019, two NIST standard indenters have been calibrated using the same set of calibration and check standards, same calibration and uncertainty procedures, but by three operators using three stylus instruments with minor changes in the calibration conditions (window size, window location, and measurement sections). The calibration results are all well within the uncertainty range, that have demonstrated the long-term calibration reproducibility of the NIST standard Rockwell hardness diamond indenters.

Keywords: Rockwell hardness, diamond indenter, microform calibration, reproducibility.

1. INTRODUCTION

Rockwell hardness (HR) is the most widely used mechanical testing method for metal products. Rockwell hardness scales are empirical, and as such are defined by the Consultative Committee on Mass and Related Quantities (CCM) – Working Group on Hardness (WGH) [1] for use by National Metrology Institutes (NMI) using the test methods specified by international standards development organizations (e.g., ISO 6508 [2,3,4] and ASTM-International E18 [5], and for use by industry. These definitions are realized at the national level through reference standards (standard testing machine and indenter) and reference testing conditions. A Rockwell hardness scale is established by the performance of a standard diamond indenter (for HRC, HRD, HRA, HR45N, HR30N and HR15N scales) using a standard testing machine and a standardized testing cycle [6]. Developments in standard hardness machines and microform calibration techniques have made it possible to establish a worldwide unified Rockwell hardness scale with metrological traceability [7]. This includes the establishment of reference standards (standard machine and diamond indenter) traceable to SI units of force, time, and length, and the establishment of reference testing conditions (a standardized common testing cycle) based on an international agreement [7].

Standard Rockwell diamond indenters play an important role for a worldwide unified Rockwell hardness scale. In 1994, the National Institutes of Standards and Technology (NIST) established a Microform Calibration System based on a stylus instrument and a set of calibration and check standards. NIST also developed calibration and uncertainty procedures for the microform calibration of Rockwell diamond indenters [8]. The expended calibration uncertainties were sufficiently low for the calibration of standard Rockwell diamond indenters [8]. In 1995, a group of 11 standard Rockwell diamond indenters were calibrated and characterized both by high geometrical uniformity and hardness performance uniformity [7]. One of them, No. 3581 was selected as the NIST primary standard for calibration of NIST's standard reference material (SRM) 2810, 2811 and 2812 Rockwell C hardness (HRC) in low, medium and high HRC range, respectively [6,9]. After about three years of calibrations with more than 3000 indentations, this indenter was recalibrated in 1997 and 2005 [10].

The No. 101 standard Rockwell diamond indenter was used for the calibration of about 200 NIST SRM Rockwell hardness blocks of the HR15N and HR30N scales [6,9]. This indenter was calibrated in 1996, 2007 and 2009 using three stylus instruments of the NIST Microform Calibration System [11]. The 1996 calibration used the original stylus instrument Form Talysurf 120L*. The 2007 calibration used a Form Talysulf PGI 1240* demo instrument before the calibration of these SRM hardness blocks. The 2009 calibration used the same type, but a different stylus instrument Form Talysulf PGI 1240 after the No. 101 indenter was used for calibrations of SRM Rockwell hardness blocks of HR15N and HR30N scales with about 2000 indentations [11].

After 2009, both indenters are barely used for regular hardness calibrations. In 2019, we re-calibrated the microform geometry of the two indenters using the standardized calibration conditions (standardized window size, window location and measurement sections) defined by the

Song, Jun-Feng; Renegar, Thomas Brian; Low, Samuel; Beauchamp, Carlos R. "Calibration Reproducibility of Microform Geometry for NIST Standard Rockwell Hardness Diamond Indenters." Presented at IMEKO 24th TC3, 14th TC5, 6th TC16 and 5th TC22 International Conference, Cavtat-Dubrovnik, HR.

recently issued ISO 6508-3:2015 standard [3] and compared the results with the previous results.

These calibration results showed high reproducibility for the microform geometry of the NIST standard Rockwell diamond indenters. It also demonstrated the high calibration reproducibility of NIST's Microform Calibration System.

In the following sections, we introduce the microform calibration system in Section 2 and the calibration and uncertainty procedures in Section 3, discuss the standardized calibration conditions in Section 4, and introduce the calibration comparison results in Section 5.

2. NIST MICROFORM CALIBRATION SYSYTEM

The NIST's Microform Calibration System was established in 1994 based on a commercial stylus instrument with an x-y-rotary stage (see Fig. 1). A set of calibration and check standards (see Fig. 2) was developed for instrument calibration and establishment of metrological traceability to the SI unit of length [8]. These standards include a calibration ball with 22 mm nominal radius, a standard wire and two ruby balls with 200 µm nominal radii, and a 120˚ angle gauge block. Three methods were used for calibrating the 2 µm nominal tip radius of the stylus instrument and for determining the value for tip radius correction [8].

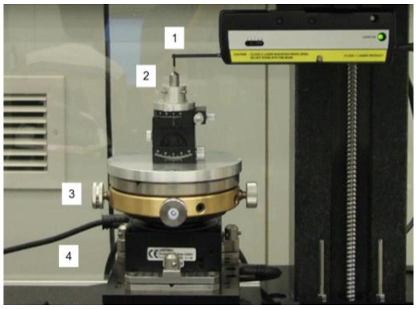


Figure 1: NIST Microform Calibration System for Rockwell diamond indenters: (1) stylus; (2) Rockwell indenter; (3) rotary stage; (4) x-y stage.

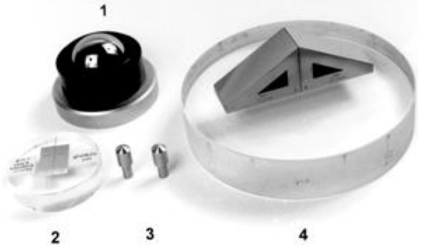


Figure 2: Calibration and check standards: (1) 22 mm radius calibration ball; (2) 200 µm radius standard wire; (3) 200 µm radius ruby balls mounted on the tip of the Rockwell indenter-shaped holders; (4) 120˚ angle gauge block.

3. CALIBRATION AND UNCERTAINTY PROCEDURE

The instrument was first calibrated with the 22 mm radius standard ball. The calibration was certified by measuring the check standard artifacts including 200 µm radius standard wire and 120⁰ angle gauge block. Then the Rockwell indenter was calibrated at nine sections with 40˚ separation. In each section, the stylus was first crowned on the top point of the diamond indenter. Then a 1.2 mm long trace was made across the crown of the indenter and 9,600 data points were collected. By windowing on the central ±90 µm part of the range and using least squares arc fitting, the least squares radius and profile deviation from the least square radius were determined. By windowing on the left and right portions of the trace, located from –460 µm to –110 µm in the left and from +110 µm to +460 µm in the right, and using the least squares line fitting algorithm, the indenter cone angle and cone flank straightness error were determined. The indenter holder axis alignment error was calculated from the cone angle measurements at nine sections. The surface roughness can also be measured using the same traced profile on the cone flank by selecting the appropriate filter cutoff. The last step in the procedure is to check the measurements by remeasuring the check standards including the standard wire and angle gauge block [8].

An uncertainty procedure was developed for calculating the calibration uncertainties of the Rockwell diamond indenters [8]. The expanded uncertainties (k = 2) of the measurement system are ±0.3 µm for the 200 µm least squares radius calibrations, and ±0.01˚ for the 120˚ cone angle calibrations. Detailed information can be found in Ref. 8. In addition, the geometric form errors of the Rockwell indenter including the profile deviations from the least square radius *Pp* and *Pv* [12], the cone flank straightness *Pt* [12], the surface roughness *Ra* [12] and the holder axis alignment error can also be calibrated [8].

4. STANDARDIZED CALIBRATION CONDITIONS

From 1995 to 2019, the calibration conditions including window size, window location and measurement sections have minor changes, partly because of there was not a clearly defined calibration conditions in the old ISO 6508 standards. In 1995 and 1996, we used eight calibration sections with 45⁰ separation, and used ±100 µm window for calibration of the least square radius and profile deviation and used ±(100 to 450) µm window for cone angle and

Song, Jun-Feng; Renegar, Thomas Brian; Low, Samuel; Beauchamp, Carlos R. "Calibration Reproducibility of Microform Geometry for NIST Standard Rockwell Hardness Diamond Indenters." Presented at IMEKO 24th TC3, 14th TC5, 6th TC16 and 5th TC22 International Conference, Cavtat-Dubrovnik, HR.

cone flank straightness calibrations [10,11]. We also used nine measurement sections with 40º separation, and ±100 µm and ± (105 to 450) µm window sizes for calibrations in 2009 [11]. The minor changes in calibration conditions may cause minor changes in calibration results, especially for extreme parameters such maximum and minimum tip radii and cone angles and profile deviations. Since the measurement sections and profile locations are not the same.

In the recently issued ISO 6508-3:2015 standard [3], the calibration conditions have been clearly defined, that makes it possible to define a standardized calibration conditions for the 2019 calibrations:

- **Measurement sections:** The ISO 6508-3:2015 standard defined "*The diamond indenter shall be measured on at least eight unique axial section planes equidistant from each other*" [3]. We use nine measurement sections with 40º separation as shown in Fig 3. The reason of not using eight sections with 45º separation is that it actually measures the same section but from opposite directions, for example, 0º and 180º, which is not unique.

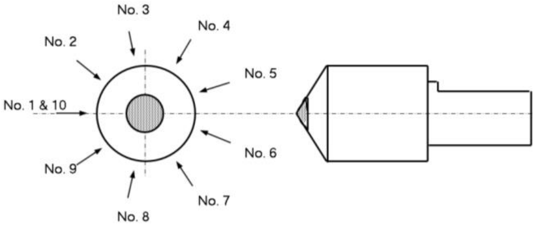


Figure 3: Nine measurement sections with 40º separation.

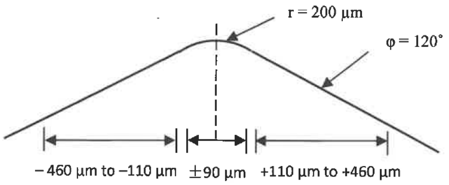


Figure 4: Window sizes and locations for calibration of Rockwell diamond indenters.

- **Window size and locations:** The ISO 6508-3:2015 standard also defined the window size and locations: "*To avoid including the blend area in the measurement of the tip radius and cone angle, the portion of the diamond surface between 90 µm and 110 µm should be ignored*" [3]. We use ±90 µm window for calibration of the least square radii and profile deviations, and use –460 µm to –110 µm in the left and +110 µm to +460 µm in the right for cone angle and cone flank straightness calibrations.

The reason to select ±(110 to 460) µm rather than ±(110 to 450) µm windows is to ensure the 0.4 mm minimum calibration length as defined by the ISO 6508-3:2015 standard: "*The mean deviation from straightness of the generatrix of the diamond cone, adjacent to the blend, shall not exceed 0,0005 mm over a minimum length of 0,4 mm*" [3].

5. CALIBRATION RESULTS

In order to maintain the long-term stability and reproducibility for the National Rockwell Hardness Scales, the calibration and check standards for the Microform Calibration System are calibrated by NIST's Dimensional Metrology Laboratory and are traceable to the SI unit of length.

The No. 3581 Rockwell diamond indenter was first calibrated in 1995 and used for the calibration of NIST SRM Rockwell C hardness blocks. In 1997, after calibrations of about 300 NIST SRM blocks with more than 3000 indentations, this indenter was recalibrated using the same Microform Calibration System. In November 2005, after the Microform Calibration System was moved to the new Nano-Metrology Center at NIST, and the calibration system was upgraded using a new type of stylus instrument, the No. 3581 indenter was re-calibrated once again [10]. The recent calibration was conducted in September 2019 using the standardized calibration conditions. These calibration results are shown in Table 1.

The variation range of the four calibrations for both the mean tip radius and cone angle from 1995 to 2019 are well within the range of the expanded measurement uncertainties (see Table 1). The profile deviations from the least square radii, including the maximum peak height *Pp*, the minimum valley depth *Pv*, and the cone flank straightness *Pt* are also without significant changes (see Table 1).

The NIST No. 101 standard Rockwell diamond indenter was used for the calibration of about 200 NIST SRM Rockwell hardness blocks of HR15N and HR30N scales [9]. This indenter was calibrated in 1996, 2007 and 2009 using three different stylus instruments of the NIST Microform Calibration System [11]. Two sets of calibration results in 2009 were from the same set of calibration data but using different window sizes of ± (100 to 450) µm and ± (105 to 450) µm for calibration of the cone angle and cone flank straightness [11]. The recent calibration was conducted in September 2019 using the standardized calibration conditions. The results of these calibrations are shown in Table 2. The variations of the calibration results for the mean tip radius and cone angle are also well within the range of the expanded measurement uncertainties.

The comparison results have demonstrated a long-term stability and reproducibility for both the microform geometry of the NIST standard Rockwell indenters, as well for that of the NIST Microform Calibration System. NIST is currently working with diamond manufacturers to develop SRM Rockwell hardness diamond indenters to support U.S. industry and international Rockwell hardness standardization.

6. SUMMARY

From 1995 to 2019, the calibration results of microform geometry for two NIST standard Rockwell hardness diamond indenters are well within the uncertainty range, that have demonstrated the long-term calibration reproducibility of NIST standard Rockwell hardness diamond indenters, as well as that of the NIST Microform Calibration System.

The long-term calibration reproducibility is mainly ensured by: 1) the calibration and check standards which are calibrated at NIST and traceable to the SI unit of length; 2) the well-established measurement techniques and uncertainty analyses; and 3) the similar calibration procedures. The calibration procedures have varied over the years due to improvements in the requirements specified by the ISO 6508 and ASTM E-18 standards [2-5]. Under such calibration and check standards, measurement techniques and procedures, the calibration results for the microform geometry of the standard Rockwell hardness diamond indenters have been highly reproducible, even when calibrated by different operators using different instruments over a 25 year span.

The early calibration during 1990's to 2000's was performed by similar conditions but with minor changes in window size, window location and measurement sections. That may cause minor variations in the calibration results. Based on the recently issued ISO 6508-3:2015 standard [3], we have used standardized calibration conditions (see Section 5) for calibration of the microform geometry of the NIST standard Rockwell hardness diamond indenters since 2019, that can further improve the calibration reproducibility in the future.

The microform geometry of the standard Rowell diamond indenters can maintain a long time period without significant changes under regular operations, even after the usage of thousands of hardness indentations. As a result, the five years calibration interval for the microform geometry calibrations of Rockwell diamond indenters as specified in the ISO 6508-3:2015 [3] standard seems very reasonable: "*Direct verification of the geometric shape shall be made before first use and at a frequency of no greater than 5 years*" [7].

Acknowledgments: The authors are grateful to Mr. A. Zheng for previous calibration of the NIST standard Rockwell hardness diamond indenters.

7. REFERENCES

[1] Website of the International Bureau of Weights and Measures (BIPM) https://www.bipm.org/wg/CCM/CCM - WGH/Allowed/International_definitions/HRC_definition.pdf

[2] ISO 6508-2:2015 Metallic Materials – Rockwell hardness test – Part 2: Verification and calibration of testing machines (scales A, B, C, D, E, F, G, H, K, N, T), ISO, Geneva, 2015.

[3] ISO 6508-3:2015 Metallic Materials – Rockwell hardness test – Part 3: Calibration of reference blocks, ISO, Geneva, 2015.

[4] ISO 6508-1:2016 Metallic Materials – Rockwell hardness test – Part 1: Test Method, ISO, Geneva 2016.

[5] ASTM E18-19 Standard Test Methods for Rockwell Hardness of Metallic Materials (West Conshohocken, PA: ASTM International), 2019.

[6] S.R. Low, *Rockwell Hardness Measurement of Metallic Materials,* NIST Special Publication 960-5, National Institute of Standards and Technology, 2001.

[7] J. Song, S.R. Low, D. Pitchure, A. Germak, S. DeSogus, T. Polzin, H. Yang, H. Ishida, and G. Barbato, "Establishing a world-wide unified Rockwell hardness scale with metrological traceability," *Metrologia*, Vol. 34(4), BIPM, Paris, pp. 331-342, 1997.

[8] J. Song, F. Rudder, T.V. Vorburger, and J. Smith, "Microform calibration uncertainties of Rockwell diamond indenters," *J. Res. NIST*, Vol. 10**0**(5), pp. 543-561, 1995.

[9] S.R. Low, R. Gettings, W. Liggett, and J. Song, "Rockwell hardness - A method-dependent standard reference material," *Proc. National Conference of Standard Laboratories*, Charlotte, NC 1999.

[10] J. Song, S.R. Low, A. Zheng, "Calibration reproducibility test for NIST No 3581 standard Rockwell diamond indenter," in Proc. XVIII IMEKO World Congress, PP. TC5-2, IMEKO, Rio de Janeiro, Brazil, 2006.

[11] J. Song, S.R. Low, A. Zheng, "Geometric measurement comparisons for Rockwell diamond indenters" in Proc. of the IMEKO XIX World Congress, Lisbon, Portugal, 2009.

[12] ASME B46.1-2012, Surface Texture (Surface Roughness, Waviness and Lay), ASME, NY, 2012.

*: 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.

Table 1: Calibration results for NIST's No. 3581 standard Rockwell hardness diamond indenter in 1995, 1997, 2005 and 2019. Uncertainties are reported for *k* = 2.

Microform Geometry Components	**Tolerances**	**Calibration Results**			
	Calibration Grade **(ISO 6508-3:2015)**	**4/11/1995**	**10/29/1997**	**11/5/2005**	**9/12/2019**
1. Spherical Radius					
1a. Mean	200 µm ±5 µm	(199.06 ±1.97) µm	(199.24 ±1.19) µm	(199.11 ±1.30) µm	(198.78 ±1.12) µm
1b. Maximum Variation	200 µm ±7 µm	Max. = 200.70 µm Min. = 197.65 µm	Max. = 201.58 µm Min. = 197.41 µm	Max. = 201.67µm Min. = 196.97µm	Max. = 201.73µm Min. = 196.57µm
1c. Profile Deviation	± 2 µm	*Pp* = 0.40 µm *Pv* = 0.29 µm	*Pp* = 0.45 µm *Pv* = 0.33 µm	*Pp* = 0.43 µm *Pv* = 0.33 µm	*Pp* = 0.40 µm *Pv* = 0.27 µm
2. Cone Angle					
2a. Mean	120° ±0.1°	119.995° ±0.020°	120.012° ±0.017°	119.987° ±0.020°	119.989° ±0.015°
2b. Maximum Variation	120° ±0.17°	Max. = 120.010°	Max. = 120.038°	Max. = 120.018°	Max. = 120.015°
2c. Cone Flank Straightness	< 0.5 µm (Mean) < 0.7 µm (Max.)	*Pt* = 0.42 µm	*Pt* = 0.49 µm	*Pt* = 0.49 µm	*Pt* = 0.38 µm
3. Holder Axis Alignment	0.3°	0.08°	0.022°	0.13°	0.13°
4. Surface Finish					
4a. Roughness Mean	---	---	*Ra* = 0.0035 µm	---	*Ra* = 0.0013 µm
4b. Max. Surface Roughness	---	---	*Ra* = 0.0036 µm	---	*Ra* = 0.0014 µm

Table 2: Calibration results for NIST standard Rockwell hardness diamond indenter No. 101 in 1996, 2007, 2009 and 2019 using three different stylus instruments. Uncertainties are reported for coverage factor, k = 2. Note: Two sets of results in 2009 are from the same calibration data using the same window size of ± 100 µm for the calibration of tip radius and profile deviations but using different window sizes of ± (100 to 450) µm and ± (105 to 450) µm for calibration of the cone angle and cone flank straightness.

Microform Geometry Parameters	**Tolerances**	**Calibration Results**				
	Calibration Grade **Specified in ISO 6508-3:2014(E)**	**4/05/1996** **Window size: ± 100 µm and ± (100 to 450) µm**	**11/02/2007** **Window size: ± 100 µm and ± (100 to 450) µm**	**4/07/2009** **Window size: ± 100 µm and ± (100 to 450) µm**	**4/07/2009** **Window size: ± 100 µm and ± (105 to 450) µm**	**9/12/2019** **Window size: ±90 µm and ± (110 to 460) µm**
1. Spherical Radius						
1a. Mean	200 µm ±5 µm	(196.83 ±0.70) µm	(196.76 ±0.72) µm	(197.09 ±0.64) µm	(197.09 ±0.64) µm	(197.16 ±0.69) µm
1b. Maximum Variation	200 µm ±7 µm	Max. = 198.25 µm	Max. = 198.15 µm	Max. = 198.15 µm	Max. = 198.15 µm	Max. = 198.70 µm
		Min. = 195.53 µm	Min. = 195.55 µm	Min. = 195.95 µm	Min. = 195.95 µm	Min. = 195.86 µm
1c. Profile Deviation	< 2 µm	*Pp* = 0.38 µm	*Pp* = 0.38 µm	*Pp* = 0.39 µm	*Pp* = 0.39 µm	*Pp* = 0.37 µm
		Pv = 0.34 µm	*Pv* = 0.31 µm	*Pv* = 0.31 µm	*Pv* = 0.31 µm	*Pv* = 0.35 µm
2. Cone Angle						
2a. Mean	120° ±0.1°	119.967° ±0.020°	119.969° ±0.021°	119.962° ±0.016°	119.955° ±0.016°	119.947° ±0.017°
2b. Maximum Variation	120° ±0.17°	Max. = 120.005°	Max. = 120.014°	Max. = 119.990°	Max. = 119.985°	Max. = 119.976°
		Min. = 119.925°	Min. = 119.936°	Min. = 119.930°	Min. = 119.925°	Min. = 119.911°
2c. Cone Flank Straightness	< 0.5 µm (Mean) < 0.7 µm (Max.)	*Pt* = 0.46 µm	*Pt* = 0.48 µm	*Pt* = 0.51 µm	*Pt* = 0.43 µm	*Pt* = 0.34 µm
3. Holder Axis Alignment	0.3°	0.092°	---	---	---	0.16°
4. Surface Finish						
4a. Roughness Mean	---	0.0043 µm	*Ra* = 0.0017 µm	*Ra* = 0.0021 µm	*Ra* = 0.0022 µm	*Ra* = 0.0013 µm
4b. Max. Surface Roughness	---	0.0051 µm	*Ra* = 0.0020 µm	*Ra* = 0.0024 µm	*Ra* = 0.0022 µm	*Ra* = 0.0015 µm

Song, Jun-Feng; Renegar, Thomas Brian; Low, Samuel; Beauchamp, Carlos R. "Calibration Reproducibility of Microform Geometry for NIST Standard Rockwell Hardness Diamond Indenters." Presented at IMEKO 24th TC3, 14th TC5, 6th TC16 and 5th TC22 International Conference, Cavtat-Dubrovnik, HR.

GAA-GAA Coexistence in the CBRS Band: Performance Evaluation of Approach 3

Weichao Gao, Anirudha Sahoo and Emma Bradford*
National Institute of Standards and Technology
Gaithersburg, Maryland, U.S.A.
Email: {weichao.gao, anirudha.sahoo}@nist.gov, emma.abe.bradford@gmail.com

***Abstract*—The General Authorized Access (GAA) users operate at the lowest priority in the Citizens Broadband Radio Service (CBRS) band. So, they must not cause harmful interference to the higher priority users and must cooperate with each other to minimize mutual interference and increase spectrum utilization. Towards this goal, the Wireless Innovation Forum (WInnForum), a standards body, has recommended three schemes. We study performance of one of the schemes, called Approach 3. To the best of our knowledge, there is no performance study available for Approach 3. WInnForum does not specify any performance metrics to evaluate the schemes. We define few performance metrics for Approach 3 that will be useful for the operators in deciding their operating parameters. We choose two actual locations and use real terrain and land cover data of continental USA in our simulation study. Hence, we expect our results to be similar to practical implementations.**

I. Introduction

The Citizens Broadband Radio Service (CBRS) band in the 3.5 GHz band has recently been opened up by the Federal Communications Commission (FCC) to the commercial operators on a priority based sharing [1]. As per the Part 96 FCC rules, there will be three tiers of users in this band. The current incumbent will operate in tier 1 with highest priority. Priority Access Lincense (PAL) users will be at the middle tier with medium priority, whereas General Authorized Access (GAA) users will be at the lowest tier operating with lowest priority. A higher tier user must be protected against harmful interference from the lower tier users. However, a lower tier user cannot expect interference protection from the higher tier users. Thus, GAA users are not protected from interference from higher tier users. The rule 47 C.F.R. § 96.35 in [1] specifies that GAA users must cooperate with each other to minimize mutual interference and increase spectrum utilization. In addition, in the first phase of deployment in the CBRS band, there will not be any PAL users. Hence, GAA-GAA coexistence plays a very important role in the commercial success of CBRS band.

The Wireless Innovation Forum (WInnForum) is the standardization body responsible for specifying the standards of various aspects of the CBRS system. The WInnForum has come out with three technical reports describing three schemes to address GAA-GAA coexistence. It is widely expected that commercial vendors will adopt one of these schemes as their GAA-GAA coexistence solution. While we have reported performance analysis of one of the schemes, called *Approach 1* [2], there is no such performance analysis available for the other two schemes. In this paper, we present our performance study of the WInnForum proposed GAA-GAA coexistence scheme called *Approach 3* [3]. It is expected that CBRS service providers will group their CBSDs into Coexistence Groups (CxGs). Each CxG will have a CxG manager which will be assigned the task of managing interference among the CBSDs belonging to the CxG.

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
AMABCC	Average Maximum Allocable Bandwidth of CBSDs in a CxG
CRC	Coverage Ratio of a CxG
RCIAC	Ratio of inter-CxG Interfered Area of a CxG
AAICIGC	Average Aggregate Interference per inter-CxG Interfered Grid of a CxG

The main contributions of this paper are as follows.

- The WInnForum technical report does not specify any performance metrics at the CxG level for Approach 3. WInnForum did not want to enforce any particular performance metrics and hence left them to the implementers and the operators. It only outlines certain principles (e.g., channel quality based on SINR), to be considered while assigning bandwidth to the individual CBSDs [3], which are pertinent to a CxG manager when it allocates BW to its CBSDs. We have defined few performance metrics at the CxG level for Approach 3 which will be helpful for operators to compare performances of CxGs in different deployment configurations.
- There is very little performance study on GAA-GAA coexistence available in the literature. In fact, to the best of our knowledge, this is the first performance evaluation

*The author was at National Institute of Standards and Technology when this work was done.

Sahoo, Anirudha; Gao, Weichao. "GAA-GAA Coexistence in the CBRS Band: Performance Evaluation of Approach 3." Presented at IEEE 45th Conference on Local Computer Networks (LCN 2020), Sydney, AU. November 16, 2020 - November 19, 2020.

of Approach 3. Hence, this work provides the very first insight into Approach 3 in terms of its performance to the research community at large.

- Our simulation study uses deployment scenarios in San Diego and Virginia Beach. These two cities are chosen because of their diverse terrain characteristics. Actual terrain and land cover data around these two cities are used by way of using the WInnForum supplied reference implementation of ITM and Hybrid propagation models [4], which use the terrain and land cover data of the continental USA. Hence, we expect our performance results to be close to practical implementations.

II. Related Work

GAA-GAA coexistence architectures in the CBRS band has been studied by the WInnForum. The WInnForum has proposed three different approaches for GAA-GAA coexistence. In these approaches, resources are allocated such that mutual interference is minimized while making sure that the allocation is fair to the CBSDs or CxGs. In Approach 1 [5], bandwidth (BW) is allocated to CBSDs such that CBSDs which potentially could interfere with each other are allocated, to the extent possible, different channels. If the deployment density is too high to achieve the above said criterion for BW allocation, then this approach would allocate the same channel to CBSDs even though they may interfere with each other. While Approach 1 considers BW as the only resource to be allocated, Approach 2 [6] considers BW and transmit power as two resources for allocation. Approach 2 tries to allocate BW at maximum allowed transmit power to CBSDs such that the CBSDs which may potentially interfere with each other are assigned non-overlapping BW. If this is not possible (e.g., due to high deployment density), transmit power is lowered to reduce the number of CBSDs which interfere with each other and then allocate non-overlapping BW to them. Approach 3 [3] tries to maximize BW allocation to CxGs using a recursive clustering approach. A CBSD, which has edges only with CBSDs belonging to the same CxG as itself, is said to belong to cluster size 1. This CBSD is allocated $100\,\%$ BW. A CBSD belonging to cluster size 2 has its edges to CBSDs which are either in its own CxG or belong to exactly one other specific CxG. CBSDs in cluster size 2 are assigned $50\,\%$ of allocable BW. This procedure is applied recursively until BW is allocated to all CBSDs. Performance analysis of Approach 1 has been reported in [2]. But this study did not consider multiple CxGs in the deployment. A simulation study of how different propagation models and deployment densities affect GAA-GAA coexistence is available in [7].

Coexistence issues in other wireless systems have also been studied. Coexistence in the shared spectrum system based on TV white space has been studied in [8]. This study includes coexistence between incumbents and secondary users as well as among secondary users. Coexistence between devices which use different air interfaces and MAC protocols but use the same spectrum has been studied. For example, the authors in [9] have studied coexistence issues between LTE-licensed assisted access (LTE-LAA) and WiFi in the 5 GHz band. In [10], [11] the authors discuss coexistence of LTE-LAA and WiFi but in the TV whitespace spectrum.

III. Overview of WInnForum Scheme: Approach 3

In this section, we give an overview of the Approach 3 [3] proposed by the WInnForum for GAA-GAA coexistence.

A. CBSD Interference Graph

Interference among CBSDs is the main concern with respect to GAA-GAA coexistence. Hence, Approach 3 starts with constructing a CBSD Interference Graph (CIG) in a given deployment area. The CIG consists of CBSDs as vertices and edges between pairs of CBSDs that interfere with each other. Since Approach 3 does not explicitly specify how to determine if an edge exists between two CBSDs, we use *area coordination* based edge creation method specified in [5]. First, we compute the coverage areas of the two CBSDs. The coverage area of a CBSD is calculated based on a propagation model, an omnidirectional antenna model and the CBSD's Equivalent Isotropically Radiated Power (EIRP). The received signal strength around the boundary of coverage is set to -96 dBm/10 MHz. Let us consider two CBSDs, C_1 and C_2. Let the coverage areas of C_1 and C_2 be A_1 and A_2 respectively. Let A be the overlap area. If the coverage areas of the two CBSDs do not overlap, i.e., $A = 0$, then there is no edge between them. Otherwise, the edge weight between them is set to $max(A/A_1, A/A_2)$. An edge is created between C_1 and C_2 if the edge weight is more than a predetermined edge threshold (ET) of the CIG. Note that once the ET is fixed for a deployment area, the CIG does not change.

1) Coexistence Groups: A CIG may consist of one or more Coexistence Groups (CxGs). A CxG consists of a group of CBSDs which manage interference among themselves. Thus, resources are allocated to individual CxGs in the CIG rather than to individual CBSDs. Each CxG typically will have a CxG manager which will then be responsible for allocating resources to individual CBSDs. In Approach 3, one or more subsets of CBSDs inside a given CxG, called clusters, are identified and are allocated BW based on the the cluster size of the subset. This information is passed onto the CxG manager. However, while allocating BW to individual CBSDs, the CxG manager may use different allocation (while not violating the cluster level allocation) to manage interference. Approach 3 essentially makes sure that BW is allocated in such a way that interference among CxGs is minimized.

2) Bandwidth Allocation: WInnForum Aproach 3 allocates BW to each CxG in a CIG. However, a subset of CBSDs (called a *cluster*) in a CxG may be allocated a certain amount of BW and another subset of CBSDs in that CxG may be allocated a different amount of BW as explained below.

Approach 3 uses a recursive clustering method to allocate BW to CxGs. It identifies a set of CBSDs in a given CxG that have edges to other CBSDs belonging to the same CxG as itself. These CBSDs are marked as belonging to cluster size 1. CBSDs in this set are allocated $100\,\%$ of GAA BW. To identify CBSDs

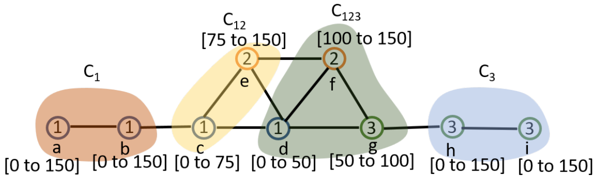


Fig. 1: An example illustrating clustering of CBSDs using Approach 3

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: Ratios of CBSD categories deployed in different area types

Area Type	Cat A	Cat B
Dense Urban	90 %	10 %
Urban	90 %	10 %
Suburban	90 %	10 %
Rural	95 %	5 %

belonging to cluster size 2, CxGs are chosen in pairs. For a pair of CxGs, CBSDs belonging to one of the two CxGs which have edges to CBSDs in its own CxG or the other CxG are identified as belonging to cluster size 2. These CBSDs are allocated 50 % of GAA BW. Similarly, CBSDs belonging to cluster size 3 are allocated 33.33 % of GAA BW and so on.

We illustrate the method followed in Approach 3 by an example shown in Figure 1. The figure shows the CIG of a CBSD deployment. Circles denote CBSDs. The numbers inside the circles denote the CxG id to which the CBSD belongs. CBSDs a and b belong to CxG1 and are connected to CBSDs which belong to CxG1 only. So, CBSDs a and b belong to cluster of size 1 (denoted as C_1). Same is the case with CBSDs h and i. Thus, they also belong to cluster of size 1 (denoted as C_3). CBSDs c and e are connected to CBSDs which belong to either CxG1 or CxG2. Hence, they belong to cluster of size 2 (denoted as C_{12}). Finally, CBSDs d, f and g are connected to CBSDs belonging to three CxGs, i.e., CxG1, CxG2 and CxG3. Hence, they belong to cluster size of 3 (denoted as C_{123}). These clusters are shown by different colors in the figure. We assume that total available bandwidth is 150 MHz. Hence, clusters C_1 and C_3 get 100 % or 150 MHz bandwidth. Thus, CBSDs a, b, h and i are assigned 0 MHz to 150 MHz. Cluster C_{12} should be assigned 50 % bandwidth. So CBSD c is assigned 0 MHz to 75 MHz whereas CBSD e is assigned 75 MHz to 150 MHz. Note that these two CBSDs are assigned 50 % bandwidth while making sure their bandwidth is non overlapping to ensure that there is no interference between CxG1 and CxG2. Finally, cluster C_{123} should be assigned 33 % bandwidth and the CBSDs belonging to CxG1, CxG2 and CxG3 in this cluster should be assigned non-overlapping frequency range. BW assigned to CBSD d should overlap with BW assigned to CBSD c (since both of them belong to the same CxG). Similarly, BW assigned to CBSD f should overlap with BW assigned to CBSD e and BW assigned to CBSD g should overlap with BW assigned to CBSD h. Thus, CBSD d is assigned 0 MHz to 50 MHz, CBSD f is assigned 100 MHz to 150 MHz and CBSD g is assigned 50 MHz to 100 MHz. It is worth pointing out that, the CBSD level frequency range allocation illustrated in this example is just one way of assigning frequencies to the CBSDs. In fact, frequency range allocation to CBSDs is a proprietary implementation by individual operators and may take into account different factors as outlined in Section 4.4 of [3].

IV. Simulation Setup

A. Deployment Model

In this study, we choose Virginia Beach (VB) in the east coast and San Diego (SD) in the west coast as our deployment locations. Two square deployment areas of $5\,km \times 5\,km$, one around VB (the center at latitude 36.842 491 and longitude -76.006 384) and the other around SD (the center at latitude 32.723 588 and longitude -117.145 319) are considered. These two locations are chosen because they have a diverse terrain. The area around VB is primarily flat, whereas it is quite hilly around SD. For ease of computation, the deployment area is divided into grids of size $50\,m \times 50\,m$. For a given deployment density, CBSDs are deployed randomly with a uniform distribution in the deployment area. Deployment density is defined as the number of CBSDs per unit area and is expressed in CBSDs per square kilometer. The relative positions of the CBSDs inside the deployment areas in SD and VB are identical. Coverage areas of CBSDs are not allowed to spill over to the outside of the square deployment area. So, if coverage of any CBSD goes outside of the square deployment area, it is clipped by the boundary of the square. CBSDs are assumed to have omnidirectional antennae. The distributions of CBSD antenna height and EIRP are shown in Table II

The total BW of CBRS band is 150 MHz. Out of this, up to 70 MHz can be used by PAL users. Hence, in this study we assume 80 MHz of BW available for GAA use. Received power threshold of −96 dBm/10 MHz is used for computation of coverage area of a CBSD, i.e., the coverage area of a CBSD is such that the received power at any point on the boundary of coverage is −96 dBm/10 MHz.

The deployment has a mix of Category A (Cat A) and Category B (Cat B) CBSDs. The ratio of Cat A and Cat B

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

CBSDs for each area type is shown in Table III. All Cat A CBSDs are considered to be indoors whereas all Cat B CBSDs are assumed to be outdoors. Note that the deployment parameters shown in the above tables are the same as those used in [2].

B. Propagation Models

We use 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] to analyze how these two models impact GAA-GAA coexistence performance. The ITM model is essentially the Longley-Rice model which is based on electromagnetic theory, terrain features and radio measurements. ITM parameters in our experiments are set as per Table IV. The Hybrid propagation model, as the name suggests, is a hybrid between ITM and extended Hata (eHata) [14] model and is proposed by the WInnForum. One limitation of the ITM model is that it does not account for clutter loss. In contrast, Hybrid model does not have this limitation. As per [13], ITM and Hybrid propagation losses are the same in the rural area. But in urban and suburban areas Hybrid propagation loss is the larger of ITM and eHata losses. Hence, the Hybrid propagation loss is generally higher than or equal to the ITM propagation loss.

C. Creation of CxGs

A set of experiments is run with a given number of CxGs. If the number of CxGs is desired to be N_g, then a CBSD is randomly placed into one of the $(N_g - 1)$ CxGs or left as a singleton CBSD. After all the CBSDs are done with placement, all singleton CBSDs are grouped into a *virtual* CxG. Note that in our experiments, relative positions of CBSDs and the grouping of CBSDs into CxGs are identical between the two deployment locations. In this study, we have set the number of CxGs to three and four in each deployment area.

D. Performance Metrics

The WInnForum does not suggest any performance metrics at the CxG level for evaluating the GAA-GAA coexistence scheme Approach 3. That task is left to the implementers and operators. In this section, we define the performance metrics which we think are appropriate and useful to the implementers and operators of a CBRS system. Traditional performance metrics such as throughput and network capacity that are used in wireless networks are not directly useful in this case. Operators in the CBRS band are more concerned about *coverage* and *interference*, since these two directly affect their deployment strategy and users' experience. Hence, the metrics we propose are based on coverage and interference and focus on quantifying the CxG-wise bandwidth allocation, inter-CxG interference, and the quality of bandwidth allocation. It is worth mentioning that the proposed metrics do indirectly affect the network capacity and throughput. The key notations used in the metrics are listed in Table V.

TABLE V: Notations used in performance metrics

CxG:	
N_g	the total number of CxGs in the deployment area
CxG_g	the CxG with index g
CBSD:	
$\mathbb{CBSD}_g$	the set of CBSDs in CxG_g
$\overline{\mathbb{CBSD}}_g$	the set of CBSDs in all the CxGs except those in CxG_g
$\mathbb{CBSD}_{k\overline{g}}$	the set of CBSDs in $\overline{\mathbb{CBSD}}_g$ that interferes with CxG_g at $grid_k$
Grid:	
$grid_k$	the grid with index k
$\mathbb{GRID}$	the set of grids in the entire deployment area
$\mathbb{GRID}_g$	the set of grids covered by the CBSDs in CxG_g
$\mathbb{GI}_g$	the set of grids in $\mathbb{GRID}_g$ that are interfered by the CBSDs from other CxGs
Power:	
rx_{ik}	received power (in dBm) at $grid_k$ from $CBSD_i$
$RX_{k\overline{g}}$	aggregate received power at $grid_k$ from all the CBSDs in $\overline{\mathbb{CBSD}}_g$
P_{th}	received power threshold used to compute coverage of a CBSD
General:	
B	total GAA BW
S_i	size of the cluster that $CBSD_i$ belongs to

1) Bandwidth Allocation:

- *Average Maximum Allocable Bandwidth of CBSDs in a CxG (AMABCC)*: To compute this metric, first Maximum Allocable Bandwidth of a CBSD (MABC) in a given CxG is computed. To compute MABC, the size of the cluster to which a CBSD belongs is computed. Let this cluster size for $CBSD_i$ be denoted as S_i, and let B be the total GAA BW. Then $MABC_i$ of $CBSD_i \in \mathbb{CBSD}_g$ is given by

$$MABC_i = \frac{B}{S_i}. \quad (1)$$

  Then the average of MABC among all CBSDs in $\mathbb{CBSD}_g$ is given by

$$AMABCC_g = \frac{\Sigma_{\forall CBSD_i \in \mathbb{CBSD}_g} MABC_i}{|\mathbb{CBSD}_g|}. \quad (2)$$

2) Inter-CxG Interference: The inter-CxG interference has two aspects: the coverage area experiencing interference and the magnitude of interference.

To define the performance metric involving coverage, we first define coverage area of CxG_g. This quantity, denoted by $GRID_g$, represents the area (in terms of grids) covered by the CBSDs belonging to CxG_g. A grid is considered covered by a CxG if the received power from any CBSD in the CxG is greater than or equal to a predefined received power threshold

P_{th} (-96 dBm/10 MHz in our deployment model). Note that the coverage of CBSDs in a CxG may overlap. Hence, $GRID_g$ is essentially the union of coverages of all the CBSDs in the CxG_g. Therefore, $GRID_g$ is given by

$$GRID_g = \bigcup_{CBSD_i \in \mathbb{CBSD}_g} \{grid_k \in \mathbb{GRID} \mid rx_{ik} \geqslant P_{th}\}. \quad (3)$$

where rx_{ik} is the received power at $grid_k$ from $CBSD_i$, and $\mathbb{GRID}$ is the set of all grids in the deployment area. Note that $GRID_g$ is independent of coexistence scheme; it only depends on the configuration of the CBSD deployment and the propagation model used to compute propagation loss.

- *Coverage Ratio of a CxG (CRC)*: This metric presents the fraction of the deployment area covered by a CxG. The CRC of CxG_g is defined as

$$CRC_g \quad = \quad \frac{|GRID_g|}{|\mathbb{GRID}|}. \quad (4)$$

  Although CRC does not directly represent interference performance of Approach 3, it could be a factor while choosing operating parameters in terms of BW and interference. Typically, operators would like to have CRC of its CxG as close to 1.0 as possible.

- *Ratio of inter-CxG Interfered Area of a CxG (RCIAC)*: To define this metric, we first need to define few intermediate terms. In a given CxG CxG_g, a grid $grid_k$ is considered to experience inter-CxG interference if there exists a pair of CBSDs ($CBSD_i$, $CBSD_j$) such that $CBSD_i \in \mathbb{CBSD}_g$ and $CBSD_j \in \overline{\mathbb{CBSD}_g}$, both the CBSDs cover $grid_k$ and the EW between them is below ET. Thus, $\mathbb{CBSD}_{k\overline{g}}$, the set of CBSDs which belong to any CxG other than CxG_g and interferes with CxG_g at $grid_k$, is given by

$$\begin{aligned}\mathbb{CBSD}_{k\overline{g}} = \{&CBSD_j \mid CBSD_i \in \mathbb{CBSD}_g, \\ &CBSD_j \in \overline{\mathbb{CBSD}_g},\, rx_{ik} \geqslant P_{th}, \\ &rx_{jk} \geqslant P_{th},\, 0 < EW_{ij} \leqslant ET\}.\end{aligned} \quad (5)$$

  Note that in the above equation, $rx_{ik} \geqslant P_{th}$ implies $CBSD_i$ covers $grid_k$, whereas $rx_{jk} \geqslant P_{th}$ implies $CBSD_j$ covers $grid_k$. When the EW between the two CBSDs is non-zero but below ET, as given by the inequality $0 < EW_{ij} \leqslant ET$, there is no edge between the two CBSDs, hence they may be assigned the same frequency range (bandwidth), which may lead to interference at the grids in the common coverage areas of the two CBSDs. Therefore, $\mathbb{GI}_g$, the set of grids in $\mathbb{GRID}_g$ which experience interference due to CBSDs from other CxGs is given by

$$\mathbb{GI}_g = \{grid_k \in \mathbb{GRID}_g \mid |\mathbb{CBSD}_{k\overline{g}}| > 0\}. \quad (6)$$

  For a given CxG, CxG_g, the $RCIAC_g$ is given by

$$RCIAC_g \quad = \quad \frac{|\mathbb{GI}_g|}{|GRID_g|}. \quad (7)$$

  Thus, RCIAC of a CxG is essentially the fraction of grids, out of all the grids covered by the CxG, that are interfered by the CBSDs from other CxGs. Closer the value of its RCIAC to zero, the better is the performance of the CxG.

- *Average Aggregate Interference per inter-CxG Interfered Grid of a CxG (AAICIGC)*: For a given CxG, CxG_g, we denote the aggregate inter-CxG interference at $grid_k$, by $RX_{k\overline{g}}$. $RX_{k\overline{g}}$ is essentially the aggregate received power at $grid_k$ from all CBSDs other than those in CxG_g. Hence, it is (in dBm) given by

$$RX_{k\overline{g}} = 10\log_{10}\big(\sum_{\forall CBSD_i \in \mathbb{CBSD}_{k\overline{g}}} 10^{rx_{ik}/10}\big). \quad (8)$$

  $AAICIGC_g$ is the average of $RX_{k\overline{g}}$ over all the grids in $\mathbb{GI}_g$ and hence is (in dBm) defined by

$$AAICIGC_g = 10\log_{10}\Big(\frac{\sum_{\forall grid_k \in \mathbb{GI}_g} 10^{RX_{k\overline{g}}/10}}{|\mathbb{GI}_g|}\Big). \quad (9)$$

  Obviously, the lower the AAICIGC of a CxG, the better is its performance.

V. Performance Results

Before we present our results, it is worth noting that Approach 3 does not prescribe a particular method for frequency range allocation at the CBSD level. Therefore, CRC values reported in our performance results are the maximum possible (i.e., upper bound) CRC values in the corresponding configurations. Similarly, RCIAC and AAICIGC values are the maximum possible (i.e., upper bound) values in the corresponding configurations.

A. Performance in terms of CRC

Figure 2 and Figure 3 show the CRC of one of the CxGs (CxG_0) in different deployment configurations when the total number of CxGs is 3 and 4 respectively. Performance of other CxGs are similar. However, they are not presented to avoid cluttering of the results. It can be observed from the figures that when density of deployment (in number of CBSDs per km^2) is low (e.g., 3), the number of CBSDs in a CxG may not be enough to cover the entire deployment area. For a given location, using Hybrid propagation model always leads to equal or lower coverage than using ITM model. This is due to higher propagation loss incurred by Hybrid model in urban and suburban areas. With respect to locations, in general, the deployment in SD tends to have less coverage than VB due to the hilly terrain which leads to more propagation loss and hence less coverage area for each CBSD. However, when deployment density is 3, the CxG in SD covers more area than in VB when ITM model is used. Recall that although the relative positions of CBSDs are same at both the locations, other parameters such as transmit power and antenna height are randomly chosen based on the area type of the location. In those cases where SD has higher CRC, there is a Cat B CBSD with large height that leads to large coverage area. This, in turn, makes the coverage of the CxG large. It can also be observed that as the number of CxGs increases from three to four, the coverage of the CxG

Sahoo, Anirudha; Gao, Weichao. "GAA-GAA Coexistence in the CBRS Band: Performance Evaluation of Approach 3." Presented at IEEE 45th Conference on Local Computer Networks (LCN 2020), Sydney, AU. November 16, 2020 - November 19, 2020.

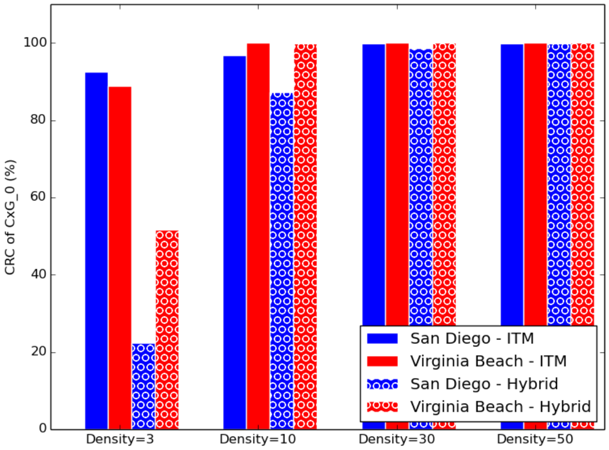


Fig. 2: Coverage Ratio of CxG_0, density in CBSD/km^2 (Number of CxGs = 3)

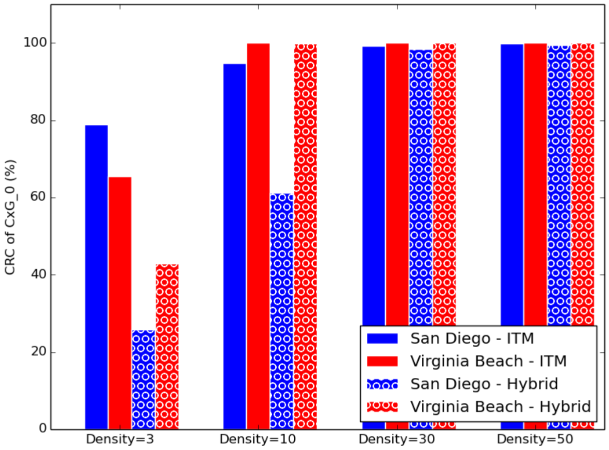


Fig. 3: Coverage Ratio of CxG_0, density in CBSD/km^2 (Number of CxGs = 4)

becomes lower when the deployment density is 3. This is due to the smaller number of CBSDs in each CxG when number of CxGs increases.

B. Performance in terms of AMABCC

Figures 4 and 5 show the AMABCC of CxG_0. It can be observed from the results that increasing the edge threshold improves the AMABCC in low density (e.g., density 3 and 10) deployment. At high densities, however, increasing edge threshold does not affect AMABCC much. At high densities, EWs are high, hence an increase in ET does not change the CIG topology much. Hence, cluster size of each CBSD does not change much and that is why AMBCC does not change much (see Eqn(2)). In terms of the propagation models, at high densities (e.g., density 30 and 50) the two models perform almost the same. At low densities Hybrid model provide slightly more flexibility in terms of increasing ET for more bandwidth, i.e., when ET increases AMABCC increase is slightly more when Hybrid model is used than when ITM is used. This is due to lower coverage when Hybrid model is used which leads to lower cluster sizes. In terms of location, deployment in SD gets more bandwidth (or AMABCC) than VB. Due to hilly terrain, overlap of coverage area of inter-CxG CBSDs is smaller at SD than VB which leads to lower cluster sizes for CBSDs and hence higher AMABCC. However, there is a deviation from this when Hybrid model is used with deployment density is 3. As discussed previously, a Cat B CBSD covers the majority of the area in San Diego making the cluster size high. This leads to lower AMABCC. Finally, as the number of CxGs increases from three to four, for a given configuration AMABCC decreases. This is because with higher number of CxGs, cluster size increases that leads to lower bandwidth. Note that the lower bound of the AMABCC occurs when all the clusters are of size N_g and is given by $1/N_g$ of the total GAA BW (80 MHz in our case), where N_g is the total number of CxGs. This is clear from Figures 4 and 5 when ET equals 0. The upper bound of AMABCC occurs when all clusters are of size one, i.e, when there is no edge between any pair of inter-CxG CBSDs. In this case, AMABCC equals the total GAA BW. This is the case when ET equals 1.0.

C. Performance in terms of RCIAC

Figures 6 and 7 show RCIAC of CxG_0 in different deployment configurations. Note that the inter-CxG interference does not exist when the ET is set to 0. It can be observed that RCIAC increases with the increase of the deployment density as well as the ET. When Hybrid model is used, RCIAC is better (smaller) than using ITM model. This is due to less overlap of CBSD coverage area caused by higher propagation loss between the inter-CxG CBSDs in Hybrid model. In terms of location, we do not see much difference in RCIAC performance between SD and VB. The hilly terrain in SD, due to higher propagation loss, produces smaller coverage area (CAC) as well as smaller interfered area compared to VB. Hence, RCIAC values at the two locations are similar.

D. Performance in terms of AAICIGC

Figures 8 and 9 show the AAICIGC performance of the CxG_0. As mentioned before, the inter-CxG interference does not exist when the edge threshold is 0. It can be observed that AAICIGC increases with the increase of ET. As ET increases, more inter-CxG edges disappear from CIG. This leads to allocating more overlapping BW to inter-CxG CBSDs which causes more interference. The deployment density does not have much impact on the AAICIGC performance. As deployment density increases, the interference power as well as the number of interfered grids increase. Thus, AAICIGC does not vary much

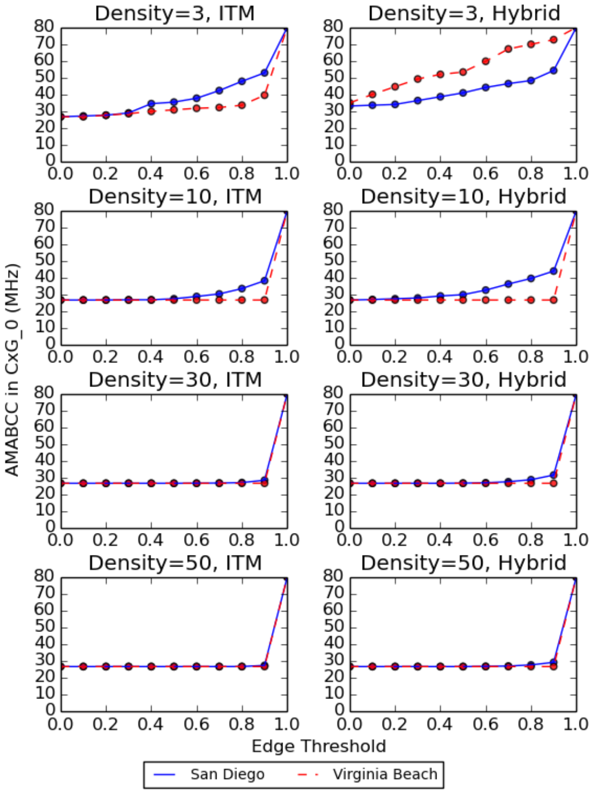


Fig. 4: AMABCC of CxG_0, density in CBSD/km^2
(Number of CxGs = 3)

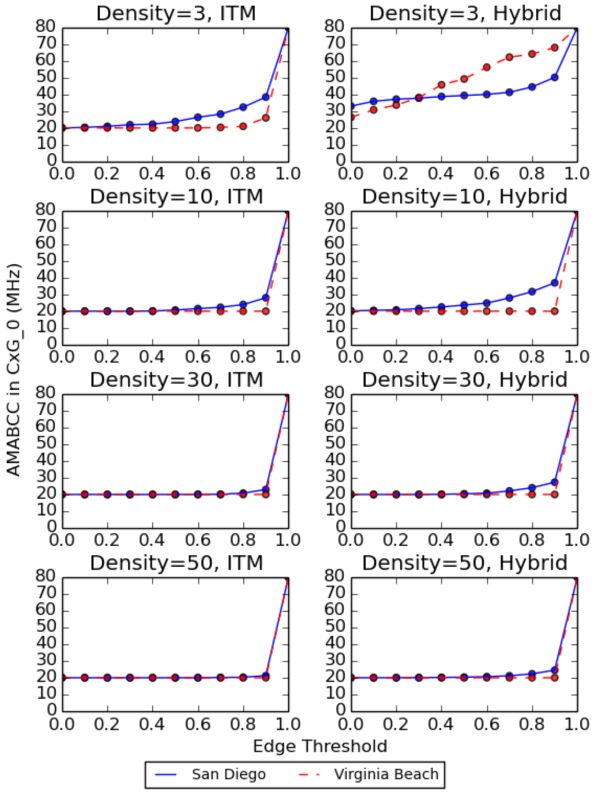


Fig. 5: AMABCC of CxG_0, density in CBSD/km^2
(Number of CxGs = 4)

(see Eqn(9)). At low ET and high deployment density (e.g., ET 0.2 and density 30), AAICIGC performance of Hybrid model is better than that of ITM model. But at high ET, ITM catches up with Hybrid. In terms of location, AAICIGC is higher in SD compared to VB when ET is small, but VB catches up when ET tends towards 1.0. This is because EWs in SD are skewed towards low values due to high propagation loss in hilly terrain. So, at low ET, more edges (which are below the ET) contribute to the interference. In contrast, EWs in VB are skewed toward high values, hence those edges contribute to interference when ET is set to high value closer to 1.0. This skewness of EWs in SD and VB is clear from the distribution of EWs shown in Figure 10.

E. *Quality of Bandwidth Allocation*

To evaluate the quality of bandwidth allocated to CxGs, one has to juxtapose the results of AMABCC and AAICIGC. At high deployment densities (.e.g, density 30 and 50), increasing ET would not produce higher AMABCC, but AAICIGC would increase rapidly. At low deployment densities, however, increasing ET does increase AMABCC, especially in SD. This comes at the cost of slight increase in AAICIGC. But low densities do not provide full coverage of the deployment area by a given CxG as is evident from the CRC performance. Therefore, these factors should be considered while choosing an operating point for a CxG (i.e., ET and deployment density).

In terms of confidence intervals for our experiments, we are able to get them for all the configurations when deployment densities are 3 and 10 $CBSDs/km^2$. When deployment density is 3, the maximum (across the two locations, two propagation models and all edge thresholds) bound of 95 $\%$ confidence interval around the mean for AMABCC is ±0.48 MHz, for RCIAC is ±0.59 $\%$ and for AAICIGC is ±1.51 dB. The corresponding numbers when deployment density is 10 are ±0.35 MHz, ±0.68 $\%$ and ±1.02 dB respectively. For higher densities, the run time to get the confidence intervals becomes prohibitively long, primarily due to large number of CBSDs in the deployment. Hence, we are not able to report confidence intervals for densities 30 and 50.

VI. Conclusion

In this paper, we evaluated performance of the WInnForum recommended GAA-GAA coexistence scheme called *Approach*

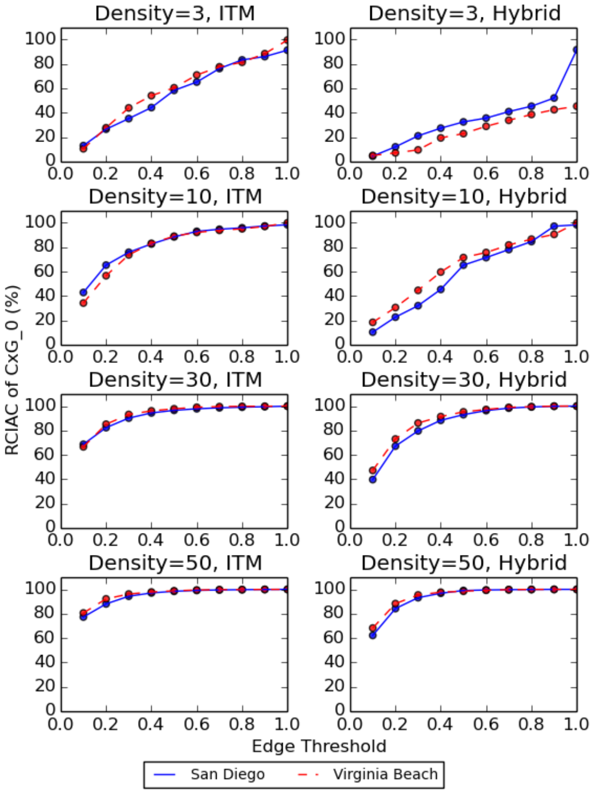


Fig. 6: RCIAC of CxG_0, density in CBSDs/km^2 (Number of CxGs = 3)

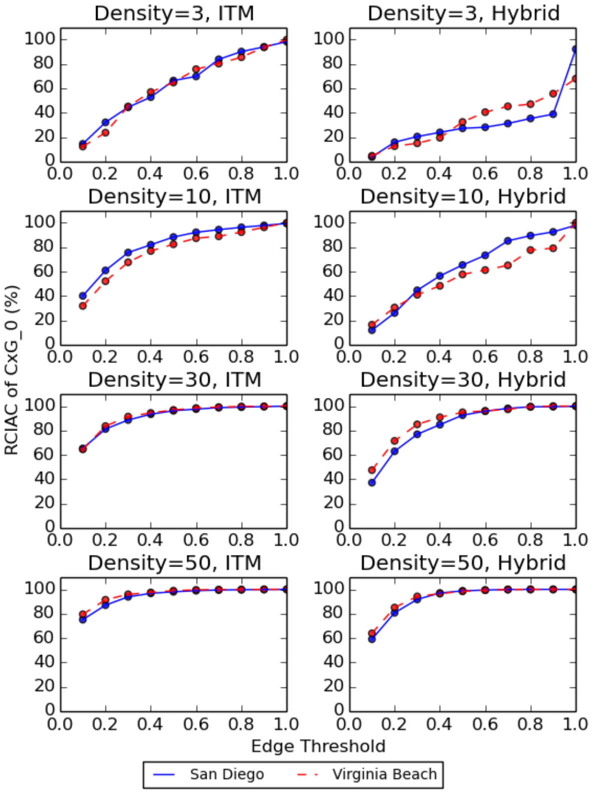


Fig. 7: RCIAC of CxG_0, density in CBSDs/km^2 (Number of CxGs = 4)

3. We looked at its performance with different deployment locations and densities using two different propagation models. If each CxG is required to cover the deployment area (most likely scenarios in metro cities), then deployment density has to be high so that CRC of the CxG is close to $100\,\%$. If CBSDs are uniform randomly distributed among the CxGs (as is done in our experiments), then at a high density deployment, when low ET is chosen, most CBSDs will belong to cluster size of N_g. Thus, a faster approximation to Approach 3 would be to just allocate $1/N_g$ of total GAA BW to each CxG. At the other extreme, if ET is close to 1.0, then an approximation is to allocate total GAA BW to each CxG.

The following are some key takeaways from our experiments. At a high density deployment, BW allocation (AMABCC) does not vary much when ET is increased, but interference increases both in terms of interference power (AAICIGC) and interference area (RCIAC) of CxGs. Hence, a good operating point is to have ET set to a low value. At a high density deployment, if getting more BW is more important, then ET should be set to a high value (e.g., greater than 0.9), but this will incur higher inter-CxG interference. On the other hand, if inter-CxG interference is of more concern than BW, then ET should be set to a low value (close to 0). In fact, ET=0 will produce no inter-CxG interference at the cost of allocating lowest BW to CxGs. These observations can be made from Figures 4 and 8. Note that, in this study we only concentrated on inter-CxG interference. There will be intra-CxG interference which needs to be managed by the respective CxG managers and is beyond the scope of this study.

References

[1] "Citizens Broadband Radio Service," 47 C.F.R. § 96, 2019.

[2] W. Gao and A. Sahoo, "Performance Study of a GAA-GAA Coexistence Scheme in the CBRS Band," in *IEEE Dynamic Spectrum Access Networks (DySPAN)*, November 2019.

[3] "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

[4] "Reference models for SAS testing." [Online]. Available: https://github.com/Wireless-Innovation-Forum/Spectrum-Access-System/tree/master/src/harness/reference_models

[5] "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

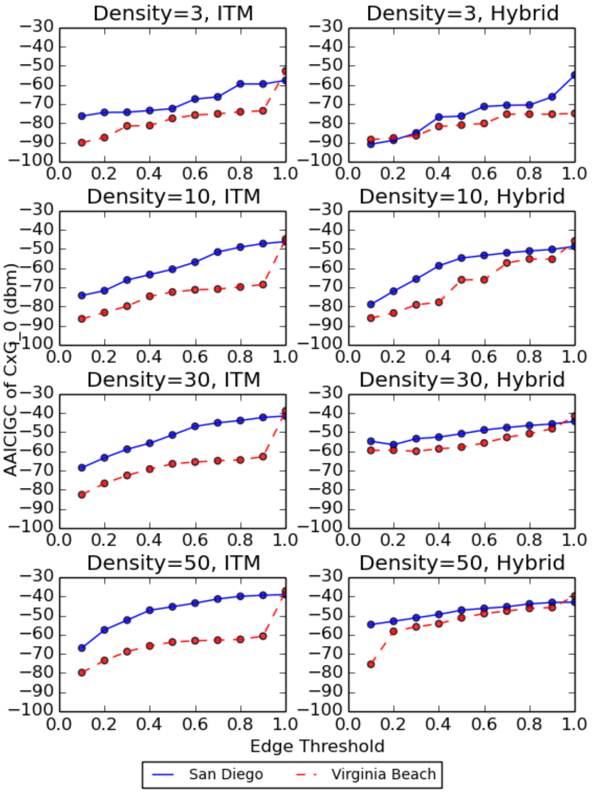


Fig. 8: AAICIGC of CxG_0, density in CBSDs/km^2 (Number of CxGs = 3)

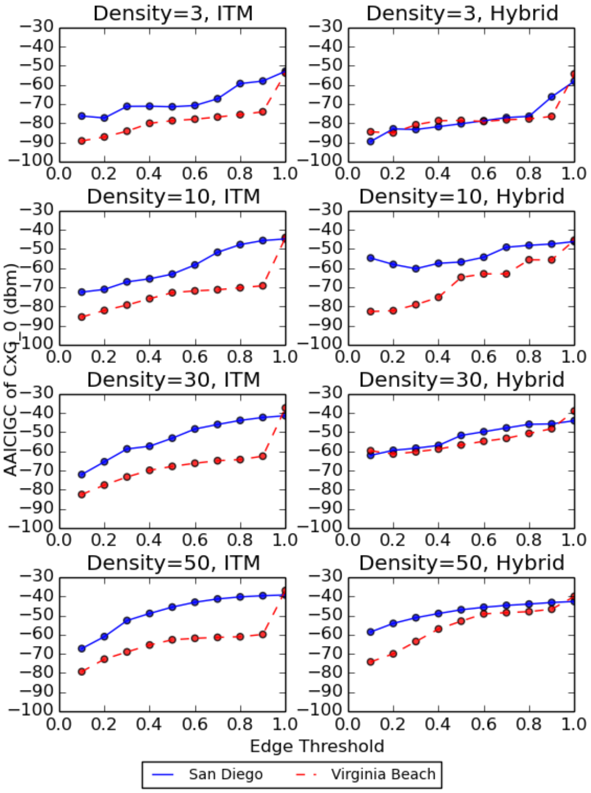


Fig. 9: AAICIGC of CxG_0, density in CBSDs/km^2 (Number of CxGs = 4)

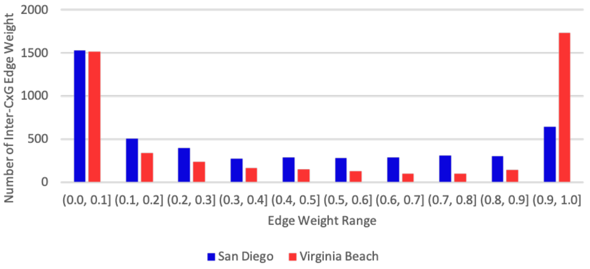


Fig. 10: Distribution of EW of all the CBSDs in CxG_0 (Density=10 CBSDs/km^2, ITM Model, CxGs=4)

[6] "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

[7] 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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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

1

An Analysis of the IOF Architecture – a Systems Integration Perspective

The Industrial Ontology Foundry (IOF) describes its ontology architecture by referencing different types of ontologies. The architecture describes these types of ontologies at a high-level. Their relationships and purposes are not clear. This research performed literature review and use case analyses with the aim to clarify the meaning of these types of ontologies. The use case analyses focused on the purpose of these types of ontologies from the perspective of enabling systems interoperability. The research was presented at the IOF face-to-face meeting on December 4, 2019 and the conclusion included in this paper is the result of discussions at the meeting.

Keywords: Industrial Ontology Foundry; Semantics; Integration; Interoperability; Types of Ontology; Ontology Architecture

1.1. Introduction

The Industrial Ontology Foundry (IOF) engages a diverse community from industry, academia, and research institutes. Its aim is to produce open-access ontologies for the manufacturing domain. The notional architecture that outlines different types of ontologies the IOF will produce and curate is shown in Figure 1-Left [IOF 20].

In this architecture, five types of ontologies are indicated including 1) Foundation Ontology (FO); 2) Domain Independent Reference Ontologies (DIROs); 3) Domain Specific Reference Ontologies (DSROs); 4) Domain Dependent Ontologies (DDOs); and 5) Application Ontologies (AOs). The IOF intends to produce and curate the first three types. The IOF charter describes these different ontology types as follows.

“The intent of these reference ontologies is to allow extensions to be progressive to more specific or constrained sub-domains (e.g., particular industry ‘verticals’ or applications). To meet this intent the IOF ontologies are expected to have an architecture that starts from alignment with a domain neutral ontology, also referred to as an Upper Ontology or **Foundational Ontology**, from which subsequent IOF ontologies

Chapter written by Boonserm Kulvatunyou*, Minchul Lee, Megan Katsumi. serm@nist.gov

Kulvatunyou, Boonserm; Lee, Minchul; Katsumi, Megan. ”An Analysis of the IOF Architecture - a Systems Integration Perspective.” Presented at 10th International Conference on Interoperability for Enterprise Systems and Applications (I-ESA 2020): Industrial Ontology Foundry (IOF) - Achieving Data Interoperability Workshop. November 17, 2020 - November 19, 2020.

can be developed (newly or adapted from existing ones) that are ontologically consistent, coherent, and modular allowing for reusability.

“Building from an FO the first ‘layer’ of IOF ontologies will be a collection of **Domain Independent Reference Ontologies** covering notions and relations independent of specific industrial domains, including time, units of measure, logistics, information, geospatial, etc. The next layer will be a collection of **Domain Specific Reference Ontologies** covering notions specific to industrial domains. Extensions following the DSROs, **Application Ontologies**, will address more focused sub-domains of industrial and manufacturing.

“The following notional diagram (Figure 1-Left) suggests how the suite of proposed IOF ontologies and their progeny may evolve. The expectation is that **the application and/or bridging ontologies** would be private or perhaps licensed.”

From the abovementioned citation, we induced that both FO and DIROs are independent of the manufacturing domain. DSROs and AOs are specific to the manufacturing domain. While FO is commonly known because only a few exist such as BFO, DOLCE, and PSL [DOL 20] [SCH 00], and there is an ISO 21838 standard [INT 20], definitions given in the charter for DIROs, DDOs and AOs are high-level or missing.

In the next section, we provide definitions of terms and notions similar to these types of ontologies. Then we perform linguistics and integration use case analysis with the aim to characterize these ontologies in more details.

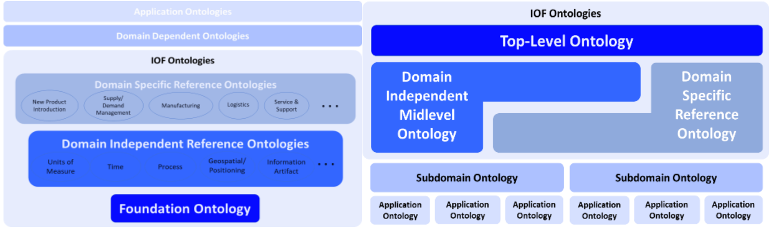


Figure. 1. *IOF ontology architecture - Left: before and Right: after – this analysis*

1.2. Literature Search

In this section, terms and definitions from literature relevant to the ontology types identified in the IOF ontology architecture are provided. Next section analyzes them in the context of the IOF terms from the practical viewpoint of systems integration.

Term	Description
Foundation Ontology	“A small, upper level ontology that is designed for use in supporting information retrieval, analysis and integration in scientific and other domains.

	BFO is a genuine upper ontology. Thus, it does not contain physical, chemical, biological or other terms which would properly fall within the coverage domains of the special sciences.” [BFO 20]
	“The primary purpose of top-level ontologies lies in providing a broad view of the world suitable for many different target domains.” [GUA 09]
Top-Level Ontology	“Ontology that is created to represent the categories that are shared across a maximally broad range of domains.” [INT 20]
Upper Ontology	Upper ontologies define top-level classes such as physical objects, activities, mereological and topological relations from which more specific classes and relations can be defined. Examples of upper ontologies are SUMO, Sowa upper ontology, Dolce, CliP, and ISO 15926-2. [BAT 05]
Upper-domain ontology	“An upper-domain ontology holds the essential core domain classes as an interface between both top-level and domain ontologies, like Organism, Tissue or Cell in the case of biology. An upper-domain ontology can also include more specific relations and further expand or restrict the applicability of relations introduced by the top-level ontology. An example for this kind of ontologies is BioTop.” [SCH 12]
Generic ontology	“Generic ontologies are valid across several domains. For example, an ontology about mereology (part-of relations) is applicable in many technical domains.” [RUD 98]
Domain Ontology	“Domain ontologies and task ontologies describe, respectively, the vocabulary related to a generic domain (like medicine, or automobiles) or a generic task or activity (like diagnosing or selling), by specializing the terms introduced in the top-level ontology.” [GUA 98]
	“Ontology whose terms represent classes or types and, optionally, certain particulars (3.3) (called ‘distinguished individuals’) in some domain” [INT 20]
	“A domain ontology includes a multitude of low-level, domain-specific classes that comprehensively describe a certain (aspect of a) domain of interest, like, Antisense RNA Transcription or DNA Replication from the Gene Ontology.” [SCH 12]
	“Domain ontologies capture the knowledge valid for a particular type of domain (e.g. electronic, medical, mechanic, digital domain).” [RUD 98]
Domain Specific Ontology	“A domain-specific ontology of concepts within a certain field, along with their relations and properties, is a new medium for the storage and propagation of specialized knowledge.” [HSI 11]
Reference Ontology (RO)	“ROs are designed to describe a certain domain adequately. They are called Reference ontologies, since they have a realist bias. Indeed, Reference ontologies contain the implicit claim that they are true about a certain portion of reality and not just that they express a more or less broad consensus among a community of experts.” [FLO 04]
	“Domain Reference ontologies represent knowledge about a particular part of the world in a way that is independent from specific objectives, through a theory of the domain.” [BUR 06]
	“ROs are an emerging ontology type that attempt to represent deep knowledge of basic science in a principled way that allows them to be re-

	used in multiple ways, just as the basic sciences are re-used in clinical applications.” [BRI 06]
	“ROs target the structuring of ontologies that are derived from them.” [GUA 09]
Application Ontology	“AOs describe concepts depending both on a particular domain and task, which are often specializations of both the related ontologies. These concepts often correspond to roles played by domain entities while performing a certain activity, like replaceable unit or spare component.” [GUA 98]
	“AOs contain all the necessary knowledge for modelling a particular domain (usually a combination of domain and method ontologies)” [RUD98]
	“AOs are suitable for direct use in reasoning engines or software packages.” [GUA 09]

1.3. Analysis

1.3.1. Foundation ontology (FO)

Foundation (or Foundational), top-level, and upper ontology are referring to the same notion. Because the ISO standard calls it Top-Level Ontology (TLO), the IOF community has agreed at the face-to-face (F2F) meeting to replace FO with **TLO** and also flip its architecture diagram upside down (Figure 1-Right).

In addition, we proposed to add to the ISO definition, from a practical ontology development viewpoint, that “TLO provides a common ground/framework to model domain ontologies across domains (e.g., common across industrial manufacturing domain and biomedical domain).” For example, when an IOF working group (WG) tried to formalize the term “Product Model”, one of the first few tasks in the formalization process was to classify the term into a category in the TLO. BFO 2.0 as the TLO, currently used by IOF, helped to establish whether the working group agreed to think of the term as a physical entity or information about it. Part of the working group argued that it should be classified as BFO’s Independent Continuant; others said they should be Generically Dependent Continuant. This resulted in a refactoring task where the “Product Model” notion was captured by both “Product” as a physical entity and “Design” as an information entity.

1.3.2. Domain Independent Reference Ontologies

We found that the term “domain independent ontology” was commonly used in literature without definition. “Generic ontology” in [RUD 98] provides a good basic definition for DIRO. However, it is not distinguishable from TLO, which is also domain independent. The key difference is only relative in that classes and properties in DIRO are subsumed by TLO’s concept, yet still applicable to multiple domains. For that reason, it was agreed at the IOF F2F meeting to change the name to “Domain Independent Mid-level Ontologies” (**DIMO**). From the ontology development perspective, IOF will have to determine whether a notion and the ontology of it shall be classified as a DIMO. For that a practical competency question is whether such notion is applicable even in a remote domain outside of the manufacturing domain, e.g., banking. If so, the term would be classified as a DIMO term. An example dilemma posted at the F2F meeting was whether the notions like "system" and "resource" should be in DIMO. Definitionally, they are widely applicable to many domains; and the IOF Core Ontology WG has specialized them into "engineered system" and "manufacturing resource" as notions specific to the manufacturing domain.

1.3.3. Domain Specific Reference Ontologies

There are three terms in the literature review that are relevant to DSRO, namely "Domain Ontology", "Domain Specific Ontology", and "Reference Ontology" (RO). Key notion of "Domain Ontology" and "Domain Specific Ontology" according to the literature is that the ontology is **low-level** and **domain-specific**. The notion captured by these two terms cannot be directly mapped to DSRO in IOF because IOF also has Domain Dependent Ontology (DDO), i.e., the notions subsume both DSRO and DDO.

The notions described in the term RO also express characteristics of DSRO. In particular, [FLO 04] stated that it is "**true about a certain portion of reality**" and a "**consensus among a community of experts**"; [BUR 06] indicated that an RO should be "**independent from specific objectives**"; and [BRI 06] said it "**represents deep knowledge**" but can be "**reused in multiple ways**". We propose a definition of DSRO, synthesized from the above three terms, as "a low-level ontology about a specific domain but still can be reused across multiple applications in the domain." Note that in the case of IOF, the domain is manufacturing. For example, a supply chain (SC) RO for manufacturing should be reusable for SC design, SC planning, and more. It should also be independent of types (e.g., push, pull), strategy, or industry. In addition, DSRO shall be subsumed by the union of TLO and DIMO.

1.3.4. Domain Dependent Ontologies

The only clue from the charter about DDO is that it is not intended to be maintained by IOF; therefore, it is quite specialized. From the linguistic analysis perspective when compared to DSRO, it is specific to a domain like DSRO, but it is not a "reference" ontology (i.e., it is not a multi-purpose ontology). Based on this analysis we proposed that the ontology be called **Subdomain Ontology** (SO) instead. Next, we analyze what the purpose of SO might be from the standard-based systems integration perspective. Since SO is not maintained by the IOF as a standard according to the current architecture, to achieve interoperability, an SO shall only include notions derived from the notions in IOF reference ontologies. In other words, SO shall be a **specialization** of **IOF ontologies** (note that this includes TLO, DSRO, and DIMO). We specifically define **specialization** using the OWL language [HIT 12] as follows. An SO is a specialization of IOF ontologies if it satisfies the following conditions 1) SO reuses some classes, properties, and axioms from IOF ontologies; 2) a new class in SO shall be a defined class with only conservative axioms[1]; 3) there shall be no new properties asserted; and 4) axioms added to classes and properties of IOF ontologies must be conservative axioms (**refinement**). The following examples[2] provide simple illustrations of two SOs called Inventory Valuation SO (IVSO) and Inventory Management SO (IMSO). IVSO imports classes Component, Inventory, Work-in-Process, Stock, Unit Cost, Sale Amount, and Purchase Amount from IOF ontologies. Inventory is refined to be a union of Work-in-Process and Stock. In addition, it defines new classes Unit Cost of Sale and Unit Cost of Purchase as a subclass of both Unit Cost and Sale Amount and a subclass of both Unit Cost and Purchase Amount, respectively. IMSO imports Component, Inventory, Unit Cost, and Work-in-Process. It, however, refines Inventory to include Stock but exclude/disjoint-with Work-in-Process. Here it can be seen that another characteristic of SOs is that they can be inconsistent with each other. We will term this horizontal inconsistency; as the inconsistency is between the same type of ontologies.

1.3.5. Application Ontologies

The definition given by [GUA 98] is well aligned with the systems integration perspective. The definition indicates that AOs describe notions for **particular tasks** and often correspond to **roles** played by domain entities while performing those tasks.

Based our experiences in systems integration, we hypothesize that AO could be an ontology that refines terms with respect to specific software application interfaces or data sources to be integrated. It refines (see refinement in 1.3.4) terms from an SO. AOs refined from the same SO could also be inconsistent. Such condition would indicate possible interoperability issues. Take, for example, a requirement for an OEM Cost Estimation Application and a Supplier PLM Application to exchange Unit Cost of Material (UCM) data. Each of them could create AOs to declare their precise semantics. The OEM AO might state that UCM is a Last Month Average cost, while Supplier AO might state that that it is a 6-Month Average cost. Ontology reasoner should be able to automatically detect such a conflict through logical inference.

1.4. Conclusion and Remark

This analysis was presented at the IOF F2F meeting at NIST on Dec 4, 2019. Attendees agreed that the research topic is essential, and that further work is needed. In the interim, this analysis resulted in a rearrangement of the architecture diagram as shown in Figure 1-Right. New questions were raised to whether more layers will be needed and whether SOs will be maintained by IOF. The community also would like to have more guidance to what kinds of axioms should be given in these different

[1] Conservative axioms are axioms containing only classes and properties from IOF ontologies

[2] The definitions of classes given here are oversimplified to convey the idea in a limited space.

Kulvatunyou, Boonserm; Lee, Minchul; Katsumi, Megan. "An Analysis of the IOF Architecture - a Systems Integration Perspective." Presented at 10th International Conference on Interoperability for Enterprise Systems and Applications (I-ESA 2020): Industrial Ontology Foundry (IOF) - Achieving Data Interoperability Workshop. November 17, 2020 - November 19, 2020.

ontologies. We planned to work with IOF WGs through additional use cases, particularly addressing the general knowledge discovery in addition to the systems integration perspective.

1.5. Disclaimer & Acknowledgement

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST. Special thanks to Evan Wallace and Melinda Hodkiewicz for their reviews and comments.

1.6. References

[IOF 20] IOF Charter, https://sites.google.com/view/industrialontologies/about/charter, 2020.

[DOL 20] DOLCE Homepage, http://www.loa.istc.cnr.it/dolce/overview.html, 2020.

[SCH 00] SCHLENOFF C., et al., The process specification language (PSL) overview and version 1.0 specification, *US Department of Commerce, National Institute of Standards and Technology*, 2000.

[INT 20] (INT 20 International Organization for Standardization. (Under development). Information technology — Top-level ontologies (TLO) (ISO/DIS Standard No. 21838). Retrieved from https://www.iso.org/standard/71954.html.

[BFO 20] BFO Homepage, http://www.ifomis.org/bfo/, 2020.

[GUA 09] GUARINO, N., et al., What Is an Ontology?, Staab S., Studer R. (eds.) *Handbook on Ontologies*. 2nd edn., p. 1–17. Springer, Heidelberg, 2009.

[SCH 12] SCHULZ, S., et al., Guideline on Developing Good Ontologies in the Biomedical Domain with Description Logics, http://www.purl.org/goodod/guideline (2012).

[RUD 98] STUDER, R., et al., Knowledge engineering: Principles and methods, *Data & Knowledge Engineering*, vol. 25 no. 1-2, 1998, p. 161-197.

[GUA 98] GUARINO, N., Formal ontology and information systems, *1st Intl. Conf. on Formal Ontology*, June 1998, Trento, Italy, IOS press, vol. 46.

[HSI 11] HSIEH, S.H., et al., Enabling the development of base domain ontology through extraction of knowledge from engineering domain handbooks, *Advanced Engineering Informatics*, vol. 25 no 2, 2011, p. 288-296.

[FLO 04] FLORIDI, L. (ed*.)., Blackwell Guide to the Philosophy of Computing and Information,* Blackwell, Oxford and New York, 2004.

[BUR 06] BURGUN, A.: Desiderata for domain reference ontologies in biomedicine. Journal of Biomedical Information vol. 39 no 3, 2006, p. 307-313.

[BRI 06] BRINKLEY, J.F., et al., A framework for using reference ontologies as a foundation for the semantic web, *Proceeding of AMIA Annual Symposium*, 2006, p. 96.

[NYA 08] NYAMSUREN, E. et al., Building Domain Independent Ontology for Web 2.0, *IEEE 8th Intl. Conf. on Computer and Information Technology Workshops*, 2008, p.655-660.

[HIT 12] HITZLER, P., et al., OWL 2 web ontology language primer. In W3C recommendation, vol. 27 no. 1, 2012, p. 123.

[BAT 05] Batres, Rafael, et al., An Upper Ontology based on ISO 15926, In: Puigjaner, L., Espuna, A. (Eds.), Proceedings of European Symposium on Computer-Aided Process Engineering, vol. 15. Elsevier, Amsterdam, p. 1543–1548.

Towards a Reference Ontology for Supply Chain Management

Farhad Ameri
Texas State University
601 University Dr.
San Marcos,TX, 78666 USA
ameri@txstate.edu

Evan Wallace
National Institute of Standards and Technology
Gaithersburg, MD, 20899, USA
Evan.wallace@nist.gov

Boonserm Kulvatanyou
National Institute of Standards and Technology
Gaithersburg, MD, 20899, USA
boonserm.kulvatunyou@nist.gov

Abstract

This paper summarizes the results from recent activities of the IOF Supply Chain Working Group (SC WG). The objectives of the IOF SC WG are to identify requirements, notions, and terms from the supply chain domain, develop a domain-specific reference ontology (DSRO), and validate the developed ontology by widening the scope to multiple use cases across the domain. The development of the current reference ontology was motivated by exploring two use cases related to supplier discovery and supply chain traceability. A draft of the OWL ontology is available for download through the provided GitHub link. This paper is not intended to provide a detailed discussion on linguistic and axiomatic analysis of the ontological entities. Rather, the intention is to provide an overview of the objectives, accomplishment, and challenges of this working group and highlight the key discussion points for the IOF workshop of I-ESA 2020

Keywords: Industrial ontologies, interoperability, supply chain

1. Introduction

As the need for supply chains, or networks, to become connected, agile, and dynamic continues, data and information integration among various supply chain participants becomes more pronounced [KIM 18]. To date, information integration is still costly, time-consuming and fragile due to the lack of *interoperability* (where we define interoperability as the ability of two or more heterogeneous, yet relevant, systems to communicate, correctly interpret, and act on information meaningfully and accurately with minimal effort [CHP 12]). The problem can be attributed to differences in the underlying semantic models and business rules implemented by different supply chain management software systems. Multiple ontologies have been proposed by researchers to resolve these differences and enable supply chain interoperability. However, the existing supply chain ontologies have failed to properly address the interoperability problem for several reasons such as weak methodological approaches, restricted and static views of supply chains, missing accounts of material traceability and service, and the dominance of taxonomies over formalized definitions [GRU 10]. To fill this gap, there is a need for a systematic ontology development methodology supported by a proper ontology architecture that includes a top-level ontology, and multiple Domain-Specific Reference Ontologies (DSRO). A DSRO can serve as shared ontology that can unify various application ontologies semantically. A proper, shared ontology architecture ensures that lower-level and application-specific ontologies are derived in a manner such that they can be used together. Such an architecture is still an open issue that will also be discussed at the IOF workshop. More details on how DSROs fill the interoperability gap can be found in [KUL 20].

Wallace, Evan K. "Towards a Reference Ontology for Supply Chain Management." Presented at 10th International Conference on Interoperability for Enterprise Systems and Applications, Tarbes, FR. November 17, 2020 - November 19, 2020.

The objectives of the IOF SC WG are to identify the requirements of a reference ontology (RO) for supply chains, develop the corresponding DSRO and other lower-level ontologies as specified by the IOF architecture [KUL 20], and validate them through multiple use cases. This paper focuses on the Supply Chain Reference Ontology (SCRO).

2. Use Cases

The development of SCRO was motivated by two industrial use cases related to supplier discovery and supply chain traceability. The ontology is intended to support the identified use cases in different ways including standardization, semantic mediation, data integration, and automated inference and reasoning. Through analyzing these use cases, two of the most important requirements for an SCRO were 1) representations of flow of materials and information and 2) characterization of organizations involved in a supply network.

2.1. Use Case Descriptions

Supplier Discovery: Supplier discovery is often a manual, slow, and inefficient process of search and requirements matching. As the interaction between suppliers and customers becomes increasingly digital, and the lifespan of SCs becomes shorter, more efficient, and intelligent; approaches to supplier search and evaluation are needed. One of the root causes of inefficiency in the process is that manufacturing companies often publish and share their capabilities using informal and unstructured representation methods. Therefore, the process is difficult to automate.

Traceability Use Case: The traceability of food products to their sources is critical for quick responses to incidents where food contamination threatens public health. Food Safety Modernization Act, a US law, now requires stakeholders in the agri-food supply chain to track food materials they acquire and sell to support timely investigation of the sources of contamination and identification of affected product. However, this has proven difficult. The causes of difficulties include diversity of stakeholders and their lexicons, systems, standards, and methods; an unwillingness to expose information of internal operations; a lack of a common understanding of the steps in a supply chain and the information needed to be collected for them; and incompleteness of data. Ontologies can be created that formally define standard Critical Tracking Event (CTE) types and associated Key Data Elements (KDE). Ontologies can also support traceability data exploration and "what if" queries to discover important relationships and fill in missing information during a traceback and trace forward effort related to a food incident.

The top terms related to the discussed use cases are listed in Table 1. Formal and subject-matter-expert definitions for a selected subset of the top terms are provided in the next section.

Table 1. IOF Supply Chain WG top-20 terms

[1] Supply Chain	[8] Production Capacity	[15] Supplier Evaluation
[2] Supplier	[9] Supplier Capability	[16] Supplier History
[3] Customer	[10] Transportation Equipment	[17] Manufacturing Service
[4] Sourcing	[11] Traceable Resource Unit-TRU	[18] Container Location
[5] Facility	[12] Tracking Event	[19] OEM
[6] Inventory	[13] Logistic Unit	[20] First-tier Supplier
[7] /Lot/Load	[14] Container/ Package/ Cargo	[21] Delivery Lead Time

2.2. Competency Questions

Competency Questions (CQ) are used to validate the ontological content against the use case requirements which is a common practice in ontology development efforts [USC 96]. Examples of competency questions related to the two use cases are provided below:

Supplier Discovery Competency Questions:

1. Which factories can machine complex geometries?
2. What is the precision machining capability of this group of companies?
3. What is the minimum wall thickness that can be machined in this factory?
4. What are the capabilities of this organization with respect to fixture design?
5. How is the performance history of this vendor with respect to on-time delivery?

Traceability Competency Questions:

1. What types of CTEs take place in the Asheville malting facility and what data is required for each?
2. Does plant 50 have all the data required for all Transfer Events that took place between 0200-1300 local time on 6 June 2018?
3. In which bins at this site was grain stored for the outbound shipment with ID 18MZ1532?
4. What Containers in the history of TRU 5384 had grain containing gluten stored in them within two weeks prior to the material in 5384 or its inputs?

3. OWL Ontology

An OWL ontology (SCRO.OWL) has been created as a pilot ontology that is based on BFO, and which uses some of the IOF Core terms [SMT 19]. The SCRO also uses classes mireotted from mid-level ontologies such as Common Core Ontology (CCO) [CCO 20]. The OWL source file is available for download through the provided link[1]. SCRO is currently being developed and extended as a single OWL file but it is likely to be partitioned into multiple modules following the modular design approach recommended by IOF technical principles. The SC Reference Ontology is intended to provide the basic ontological constructs needed to represent both the structure (supply chain members and their roles, functions, and relations) and the operation (processes and flow of material and information) of industrial supply chains. There are two central notions in SCRO: 1) Group of Suppliers and 2) Supply Chain System. A *Group of Suppliers* is a group of agents (person or organization) who play causal roles in manufacturing products or providing services in the context of a specific supply chain. A Supply Chain System, on the other hand, is an *Engineered System* comprising all agents, equipment, facilities, software systems, and other systems and resources, governed by a set of rules, designed and deployed with the function of delivering a product or a service to some customer. Accordingly, Group of Suppliers is part of a Supply Chain System. In the domain of supply chain management, the term ‘supply chain’ is the generic term often used to refer to both the group of organizations (that participate in a supply chain) and the supply chain system. To avoid confusion, unambiguous labels are selected for these two closely related notions. Figure 1 shows the class diagram for some of the core classes and relationships in the SCRO.

[1] https://github.com/InfoneerTXST/SupplyChainOntology

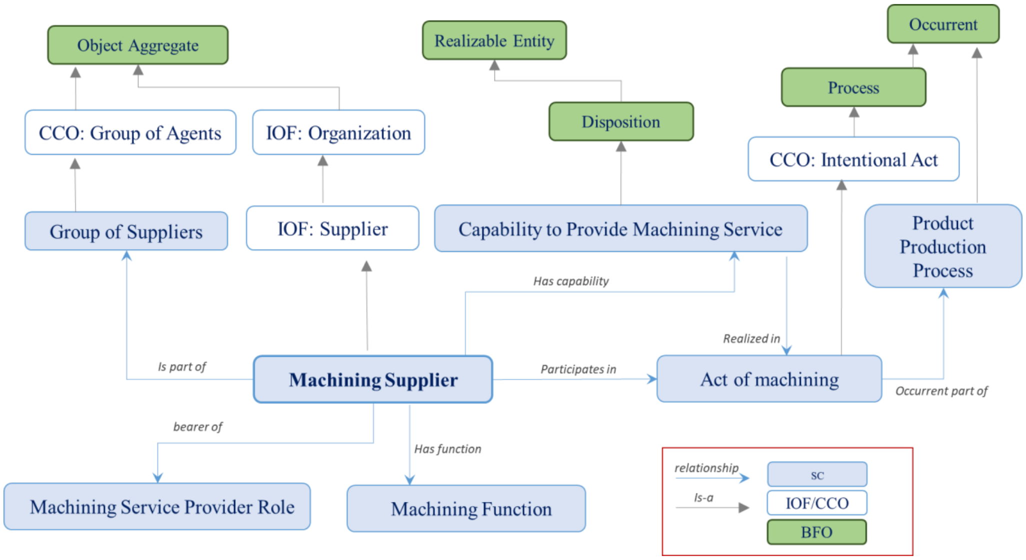


Figure 1. The class diagram for some of the core classes of the supply chain reference ontology

The classes and properties that are included in the SCRO's early draft are mainly geared towards describing the agents that participate in the supply chain and their roles, functions, and capabilities. There are two main types of Supply Chain Role included in SCRO, namely, Product Supplier Role and Service Provider Role. Manufacture Role, Wholesaler Role, and Distributor Role are example sub-classes of Product Supplier Role. Test Service Provider Role and Transportation Service Provider Role are examples of supply chain service provider role. Those roles can be inhered in various supply chain agents regardless of their nature of business.

SCRO is also intended to provide the ontological constructs for formal modeling and representation of organizational capabilities since they are crucial in supply chain design and planning phase. There are two possible approaches for capability representation in SCRO. The first approach (Approach A in Figure 2) is to use Modal Relations Ontology (MRO) to represent the processes and services a potential supplier can provide in future. The second approach (approach B) is to directly assert the capability instances for a given supplier. Those instances of capability can be realized in future processes that the supplier will participate in them once selected as a member of some supply chain.

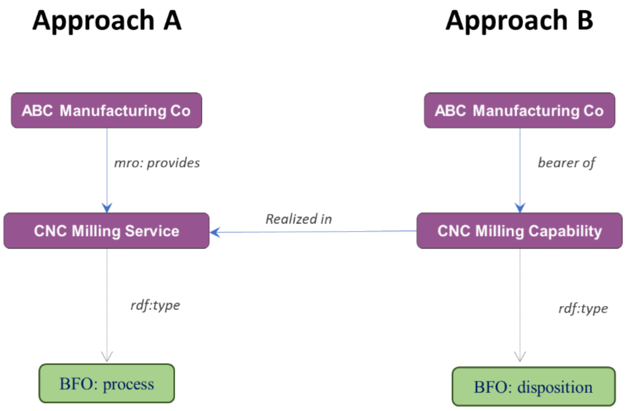


Figure 2. Two approaches for representing the capabilities of a manufacturing company

Wallace, Evan K. "Towards a Reference Ontology for Supply Chain Management." Presented at 10th International Conference on Interoperability for Enterprise Systems and Applications, Tarbes, FR. November 17, 2020 - November 19, 2020.

The current draft ontology is not fully axiomatized yet since the initial focus was on providing accurate natural language (NL) definitions. Table 2 provides the Natural Language (NL) and Semi-formal NL definitions for some core notions of the ontology. Semi-Formal NL is a human friendly version of the FOL axiom. The FOL axioms for some of the more stable classes are provided in Table 2 as well.

Table 2. Definitions and axioms for a selected subset of terms

Supplier	**NL Def.**	An organization or person who sells products or services.
	Semi-Formal	An Agent who bears a Supplier Role.
	FOL Axiom	*Instance-of(x, supplier,t) ≡ instance-of(x,agent,t)& ∃ y(supplier-role(y) & has-role(x,y,t)).*
Supplier Role	**NL Def.**	"role" classes are not user-facing. Therefore, no SME definition is provided for them.
	Semi-Formal	A Role inhering in an Agent that, if realized, is realized in some act of selling.
	FOL Axiom	*supplier-role(x) ≡ ∃y(agent(y) & has-role(y, x) & ∀ p(process(p) & realizes(y, p)) → act-of-selling(p))*
Group of Suppliers	**NL Def.**	Supply chain is a set of companies and other organizations involved in trading and other business relationships with one another
	Semi-Formal	A Group of Agents who are parts of some Supply Chain System and play causal role (are agent) in some Product Production Process that outputs some Product or in some Service Provisioning Process that outputs some Service.
	FOL Axiom	*instance-of (x, supply-chain, t) ≡ instance-of(x,group-of-agents,t)& ∀y(member-of(y,x)& ∀process(p) &participate-in(y,p)) &occurrent-part-of(p,scp) & instance-of(scp, supply-chain-product-production-process)or instance-of(scp, supply-chain-service-provisioning-process)*
Mfg. Service	**NL Def.**	A valuable action performed to satisfy a need or to fulfill a demand related to manufacturing a product.
	Semi-Formal	A Planned Process in which a supplier performs a manufacturing process for a customer and in which service provisioning and consumption occur within the same temporal region.
	FOL Axiom	*Instance-of (x, manufacturing service, t)→ Instance-of (x, planned process, t)*
Mfg. Capability	**NL Def.**	The ability of a resource (such as a human agent, an organization, or an equipment) to achieve some desired manufacturing outcome usually through employing some additional resources
	Semi-Formal	A disposition whose realizations brings benefits to an agent or group of agents and can be graded on a scale from zero to positive.
Sourcing	**NL Def.**	The process of identifying a company that provides a needed good or service.
	Semi-Formal	A planned process with the specified output of an identified supplier who can provide a service or a product.

4. Conclusion and Next Steps

This paper reports the work in progress by the IOF towards creating a reference ontology for the supply chain domain. The focus of the Supply Chain WG in the first phase has primarily been on representing the *continuant* side of the supply chain domain. In the next phase, the supply chain processes will be formalized. The notion of Service, and its sub-types including Manufacturing Service, also needs further formalization and axiomatization. The taxonomy of supply chain roles also needs further expansion. Along the way, several ontological challenges were encountered, and still need to be addressed. For example, it is not yet verified if using the Process Aggregate class is the right approach for representing the collection of processes that occurs in a supply chain. In modeling the notions based on requirements from the traceability use case, the ontological quandaries include providing a better means for constructing the *history* of supply chain that can capture the flow of materials across various geospatial and temporal regions. Furthermore, representing supplier capabilities also poses a host of challenges since capability is a complex and multi-faceted notion. Ontology modularization also needs to be addressed in a more systematic fashion in IOF. Currently, SCRO selectively imports classes from CCO that are of direct interest and relevance for the supply chain uses cases in hand. As a group, the IOF should decide whether CCO, IAO, and other mid-level ontologies should be imported as a whole or class mireotting is an acceptable practice.

5. References

[KIM 18] KIM, H. M., and LASKOWSKI, M., 2018, "Toward an ontology-driven blockchain design for supply-chain provenance," *Intelligent Systems in Accounting Finance & Management*, 25(1), pp. 18-27.

[CHP 12] CHAPURLAT, V., and DACLIN, N., 2012, "System interoperability: definition and proposition of interface model in MBSE Context," *IFAC Proc.* Vol. 45(6), pp. 1523-1528.

[GRU 10] GRUBIC, T., and FAN, I.-S., 2010, "Supply chain ontology: Review, analysis and synthesis," *Computers in Industry*, 61(- 8), pp. 776-786.

[ARP 15] ARP, R., SMITH, B., & SPEAR, A. D. “Building ontologies with basic formal ontology”. MIT Press, 2015.

[SMT 19] SMITH, B. A First-Order Logic Formalization for the Industrial Ontology Foundry Signature Using Basic Formal Ontology, *Joint Ontology Workshop (JOWO), 10th International Workshop of Formal Ontology Meet Industry (FOMI),* Graz, Austria, September 23-25, 2019.

[CCO 20] CUBRC, 2020. “Common Core Ontology,” https://github.com/CommonCoreOntology/CommonCoreOntologies.

[KUL 20] KULVATUNYOU, B., Lee, M. and Katsumi, M., 2020 “An Analysis of IOF Architecture, A Systems Integration Perspective,” I-ESA IOF Workshop, In: Enterprise Interoperability IX. Proceedings of the I-ESA Conferences, vol 10. Springer.

[RUD 20] Rudnicki, R., 2020. “An Overview of the Common Core Ontologies,” Available at: https://github.com/CommonCoreOntology/ [Online; accessed 4 Nov 2020].

[USC 96] Uschold, M., Gruninger, M. “Ontologies: principles, methods, and applications,” Knowledge Engineering Review, 11 (2) (1996), pp. 93-155.

A Historical and Statistical Study of the Software Vulnerability Landscape

Assane Gueye
Carnegie Mellon
University Africa
Kigali, Rwanda
assaneg@andrew.cmu.edu

Peter Mell
National Institute
of Standards and Technology
Gaithersburg MD, USA
peter.mell@nist.gov

***Abstract*—Understanding the landscape of software vulnerabilities is key for developing effective security solutions. Fortunately, the evaluation of vulnerability databases that use a framework for communicating vulnerability attributes and their severity scores, such as the Common Vulnerability Scoring System (CVSS), can help shed light on the nature of publicly published vulnerabilities. In this paper, we characterize the software vulnerability landscape by performing a historical and statistical analysis of CVSS vulnerability metrics over the period of 2005 to 2019 through using data from the National Vulnerability Database. We conduct three studies analyzing the following: the distribution of CVSS scores (both empirical and theoretical), the distribution of CVSS metric values and how vulnerability characteristics change over time, and the relative rankings of the most frequent metric value over time. Our resulting analysis shows that the vulnerability threat landscape has been dominated by only a few vulnerability types and has changed little during the time period of the study. The overwhelming majority of vulnerabilities are exploitable over the network. The complexity to successfully exploit these vulnerabilities is dominantly low; very little authentication to the target victim is necessary for a successful attack. And most of the flaws require very limited interaction with users. However, on the positive side, the damage of these vulnerabilities is mostly confined within the security scope of the impacted components. A discussion of lessons that could be learned from this analysis is presented.**

***Index Terms*—Vulnerabilities, Statistics**

I. Introduction

Understanding the landscape of software vulnerabilities is a key step for developing effective security solutions. It is difficult to counter a threat that is not well understood. Fortunately, there exist vulnerability databases that can be analyzed to help shed light on the nature of publicly published software vulnerabilities. The National Vulnerability Database (NVD) [1] is one such repository. NVD catalogs publicly disclosed vulnerabilities and provides an analysis of their attributes and severity scores using the Common Vulnerability Scoring System (CVSS) [2]. CVSS is used extensively by security tools and databases and is maintained by the international Forum of Incident Response and Security Teams (FIRST) [3].

The CVSS provides a framework for describing vulnerability attributes and then scoring them as to their projected severity. The attributes are metric values that are the input to a CVSS equation that generates the score. It is the vulnerability attribute descriptions (the metric values) that are of primary interest to our work, although we also look at the raw scores. The use of CVSS by vulnerability databases provides a suite of low level metrics, encapsulated in a vector, describing the characteristics of each vulnerability. CVSS was initially released in 2005 [4], was completely revamped with version 2 (v2) in 2007 [5], and was updated with new and modified metrics in 2015 with the release of version 3 (v3) [6]. Note that a minor update version 3.1 was released in 2019 [7], but the changes therein do not affect our work. The software flaw vulnerability landscape was thoroughly analyzed in the scientific literature using v2 when it was first released [4], [8]–[13], but little work has been done since to evaluate changes to that landscape over time. Also in our literature survey, we did not find a single study that uses the updated and significantly modified v3 to understand the software vulnerability landscape.

In this paper, we use the CVSS v2 and v3 data provided by the NVD to undertake a historical and statistical analysis of the software vulnerabilities landscape over the period 2005 to 2019. More precisely, we conduct three studies analyzing the following:

- score distributions,
- metric value distributions,
- and relative rankings of the most frequent metric values.

For our first study, we analyze and compare the distributions of CVSS v2 and v3 scores as generated from the NVD data. We then compare the empirical distributions against the theoretical score distributions, assuming that all CVSS vectors are equally likely (which is not the case, but it is illustrative to evaluate the differences).

For our second study, we compute the distributions of the CVSS metric values (i.e., vulnerability characteristics) for each year. We then analyze the differences from 2005 to 2019 to determine if and how vulnerability characteristics change over time.

For our third study, we identify the most frequent metric values and analyze their relative rankings from 2015 to 2019. For each year and for both CVSS versions, we compute the values of the top 10 observed vulnerability metrics as well as their frequencies. We then generate parallel coordinates plots showing the values and frequencies of each metric for each year.

Gueye, Assane; Mell, Peter. "A Historical and Statistical Study of the Software Vulnerability Landscape." Presented at Seventh International Conference on Advances and Trends in Software Engineering (SOFTENG 2021), Porto, PT. April 18, 2021 - April 22, 2021.

Our analysis shows that the software vulnerability landscape has been dominated by only a few vulnerability types and has changed very little from 2005 to 2019. For example, the overwhelming majority of vulnerabilities are exploitable over the network (i.e., remotely). The complexity to successfully exploit these vulnerabilities is dominantly low while attackers are generally not required to have any level of prior access to their targets (i.e., having successfully authenticated) in order to launch an attack. And most of the flaws require very limited interaction with users. On the positive side, the damage of these vulnerabilities is mostly confined within the security scope of the impacted components. Few vulnerabilities obtain greater privileges than are available to the exploited vulnerable component.

Our findings are consistent with previous studies [8] (mainly based on CVSS version 2). This indicates that the same vulnerabilities are still being found in our software, suggesting that the community has not been doing a great job correcting the most common vulnerabilities.

The remainder of this paper is organized as follows. Section II presents the CVSS data sets that constitute the basis of our study. Section III gives the details of our analysis and our discussion. Section IV provides a summary of related work and Section V concludes.

II. The CVSS Datasets

CVSS consists of three metric groups: base, temporal, and environmental. The base group represents the intrinsic qualities of a vulnerability that are constant over time and across user environments, the temporal group reflects the characteristics of a vulnerability that change over time, and the environmental group represents the characteristics of a vulnerability that are unique to a user's environment [6]. In this work, we evaluate only the base metrics as no extensive database of temporal scores exists and the environment metrics are designed for an organization to customize base and temporal scores to their particular environment.

Tables I and II show the base score metrics and possible values for v2 and v3, respectively. The CVSS base score takes into account the exploitability (how easy it is to use the vulnerability in an attack) and impact (how much damage the vulnerability can cause to an affected component) of a vulnerability apart from any specific environment.

The exploitability score is determined by the following:

- attack vector (v2 & v3): 'the context by which vulnerability exploitation is possible',
- attack complexity (v2 & v3): 'the conditions beyond the attacker's control that must exist in order to exploit the vulnerability',
- authentication (v2): 'number of times an attacker must authenticate to a target in order to exploit a vulnerability',
- privileges required (v3): 'the level of privileges an attacker must possess before successfully exploiting the vulnerability', and
- user interaction (v3): a human victim must participate for the vulnerability to be exploited.

TABLE I
CVSS v2 metrics

CVSS v2 Metrics	Metric Values
Access Vector (AV)	Network (N), Adjacent (A), Local (L)
Attack Complexity (AC)	Low (L), Medium (M), High (H)
Authentication (Au)	Multiple (M), Single (S), None (N)
Confidentiality (C)	Complete (C), Partial (P), None (N)
Integrity (I)	Complete (C), Partial (P), None (N)
Availability (A)	Complete (C), Partial (P), None (N)

TABLE II
CVSS v3 metrics

CVSS v3 Metrics	Metric Values
Attack Vector (AV)	Network (N), Adjacent (A), Local (L), Physical (P)
Attack Complexity (AC)	Low (L), High (H)
Privileges Required (PR)	None (N), Low (L), High (H)
User Interaction (UI)	None (N), Required (R)
Scope (S)	Unchanged (U), Changed (C)
Confidentiality (C)	High (H), Low (L), None (N)
Integrity (I)	High (H), Low (L), None (N)
Availability (A)	High (H), Low (L), None (N)

The impact score (v2 & v3) is determined by measuring the impact to the confidentiality, integrity, and availability of the affected system. Also included (v3) is a scope metric that 'captures whether a vulnerability in one vulnerable component impacts resources in components beyond its security scope'.

A particular assignment of metric values is then used as input to the CVSS base score equations to generate scores representing the inherent severity of a vulnerability in general, apart from any particular environment. The raw score in the range from 0 to 10 is then often translated into a 'qualitative severity rating scale' (None: 0.0, Low: 0.1 to 3.9, Medium: 4.0 to 6.9, High: 7.0 to 8.9, and Critical: 9.0 to 10.0) [6].

Vulnerability analysts apply the metrics to vulnerabilities to generate CVSS vector strings. The vectors describe the metric values, but not the CVSS scores, for a particular vulnerability using a simplified notation.

The NVD is the 'U.S. government repository of standards based vulnerability management data' [1]. It provides CVSS vectors and base scores for all vulnerabilities listed in the Common Vulnerabilities and Exposures (CVE) [14] [15] catalog of publicly disclosed software flaws. We use NVD to evaluate both CVSS v2 and v3 vectors and scores. The v2 data covers all CVE vulnerabilities published between 2005 and 2019. The v3 data ranges from 2015 to 2019 (only limited v3 data is available prior to 2015). These coverage dates result in the inclusion in our study of 118 173 v2 vectors and scores and 55 441 v3 vectors and scores.

III. Data Analysis

We analyze the NVD CVSS data in order to better understand the software vulnerability landscape. We investigate both the current nature of the threat posed by the existence and public disclosure of these vulnerabilities as well as how this threat has changed over time. To achieve this, we conduct the three studies described previously where we analyze the following: score distributions, metric value distributions, and relative rankings of the most frequent metric values.

A. Score Distributions

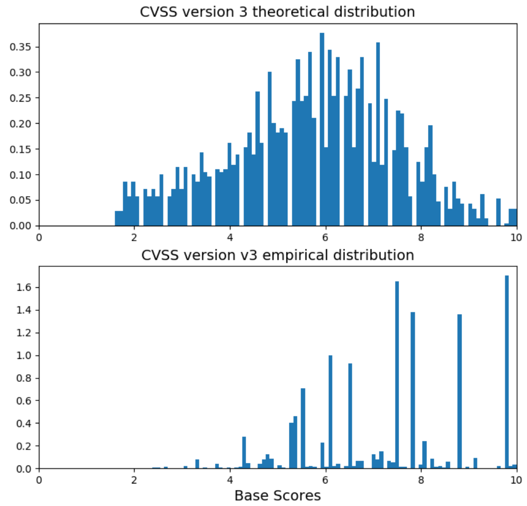


Fig. 1. Theoretical vs Empirical Score Distributions for CVSS version 3. The y-axis shows the numerical values of the base scores of vulnerabilities. The top figure is obtained by considering all possible assignments of metric values, while the bottom figure corresponds to scores of actual vulnerabilities discovered in software.

The top graph of Figure 1 shows the theoretical distribution of the v3 scores (v2 scores are similar and not shown in the paper due to space limitation. They can be found in the appendix of [16]). These plots show what is expected if all CVSS vectors (i.e., vulnerability types) are equally likely to occur. Note how the theoretical distribution was designed, by the FIRST CVSS committee, to spread CVSS scores throughout the range with a somewhat normal distribution with the most probable scores occurring in the middle of the distribution (a little biased to the right). That said, it is interesting in that for both v2 and v3 some scores are not possible even though they lie within the valid range of score values.

The empirical distribution is shown in the bottom of Figure 1 for v3. The empirical data indicates a predominance of certain vectors (groupings of vulnerability characteristics) in the real world. Thus, only a few vulnerability feature sets describe the majority of publicly disclosed vulnerabilities. This leads to the frequent use of just a very small number of scores. A similar observation was made in a previous study of the v2 scoring system [8].

The results observed with v3, which uses data from 2015 to 2019 (since v3 vectors are not generally available prior to 2015) are similar to those with v2, which uses data from 2005 to 2019. Hence, the long-term obtained with CVSS v2 data is confirmed by the shorter-term data of CVSS v3.

B. Metric Value Distributions

To investigate more carefully (in order to identify) possible differences per year and trends over time, we focus on the distributions of each set of metric values per year over the time period of study. Figure 2 provides the histograms for v3 from 2015 to 2019. We have also plotted the histograms for v2 [16], which cover from 2005 to 2019. The inclusion of v2 in the study allows for a comparison over 15 years as opposed to being limited to just 5 years with v3, due to its more recent development.

The histograms for individual metric values for v3 appear almost the same year to year for the 5 years of study. This is the same in v2 over the longer time period of 15 years with some small exceptions: in 2014, the attack vector (AV) value of adjacent had some significance. According to the NVD team [17], this was a one time anomaly due to more than 800 CVEs all being announced simultaneously by an organization doing analyses on phone apps. The Attack Complexity (AC) value ‘Medium’ increased some from 2007 onward, but then was steady, the Authentication (Au) value ‘Single’ increased slightly over the years, and the Confidentiality (C), Integrity (I), and Availability (A) metric proportions between ‘None’, ‘Partial’, and ‘Complete’ varied slightly from year to year while generally maintaining themselves about the same.

Overall though, the software vulnerability landscape for publicly disclosed vulnerabilities has been almost static during the period of study. This said, doing comparisons between the v2 and v3 histograms, we see some differences, but this is due to differences in the approaches of the two versions of CVSS. These differences are primarily seen in regards to the metrics C, I, and A, which we will discuss shortly.

Consider the AV metric which reflects the context by which the vulnerability can possibly be exploited: Network (N), Adjacent (A), Local (L), or Physical (P). Both data sets show a high peak at N, a low peak at L and almost nothing at A and P. This indicates that the overwhelming majority of publicly disclosed software vulnerabilities are exploitable over the network (i.e., remotely), and it has been that way consistently through the period of study.

The AC metric describes the conditions beyond the attacker’s control that must exist in order to exploit the vulnerability. When it is low (AC:L), the attacker can expect repeatable easy successes, while when it is high (AC:H) the attack is less likely to be successful. The data shows that the AC metric is largely dominated by the values of AC:L for v3 and AC:L and AC medium (AC:M) for v2. This indicates that the set of publicly disclosed vulnerabilities have been predominantly easy to exploit.

This “easiness” to exploit vulnerabilities is confirmed by the other metrics for each CVSS version. For v3, the Privileges Required (PR) metric describes the level of privileges an attacker must possess before successfully exploiting a vulnerability. The User Interaction (UI) metric captures the requirements for a human user (other than the attacker) to participate in the successful compromising of the vulnerable

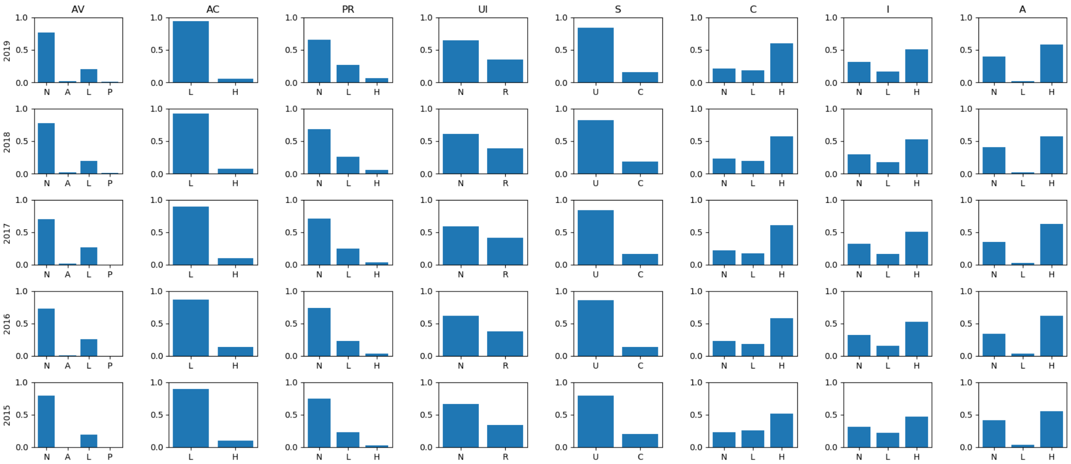


Fig. 2. CVSS v3 metrics' values distributions over the years

components. The data shows that in most of the cases, no privilege is required and very little user interaction is needed for a successful attack.

Similarly, with v2, the Au metric measures the number of times an attacker must legitimately authenticate to a target in order to be in a position to exploit a vulnerability. The data shows that almost always, there is no authentication required prior to exploiting a vulnerability. Sometimes a single authentication is required, but almost never is there a vulnerability that requires multiple authentications in order to be successfully exploited.

CVSS v3 introduced a new Scope (S) metric, which captures the spill-over effect: how much a vulnerability in one vulnerable component impacts resources in components outside of its security scope. When the scope is unchanged (S:U), there is no spill-over, while when the scope is changed (S:C) the vulnerability will very likely affect other components. The data shows that the scope metric has predominantly been S:U.

The last three metrics C, I, and A are common to both CVSS versions. They capture the extent to which a successful exploitation of a vulnerability will affect these three principles of security on the effected component. With respect to these metrics, the v3 data shows that the impact on C, I, and A has predominantly been high (C:H, I:H, and A:H) with very similar distributions for all the years. The v2 data also shows a similar stationary behavior in the distributions. However, the difference in the fraction of high for v3 and complete for v2 is notable. One might expect the high values in CVSS v3 to be equivalent to the complete values for v2. However, this is not the case as they are defined differently. According to the NVD team [17] "the CVSS scoring systems are fundamentally different regarding qualifications for CIA Partial/Complete and Low/High. This is a common misconception due to the nuances of the scoring systems. In addition to this, the NVD takes the approach of representing the worst-case scenario when information is lacking. This typically results in default values of HIGH being attributed to a CVE unless data is available that identifies a limitation to the impact or meets qualifying text for the specification."

C. Relative Rankings of the Most Frequent Metric Values

We now focus on the most prominent individual values of the metrics, evaluating the rankings of the top 10 metric values observed each year and providing a comparison between the years. Figure 3 shows the rankings for v3 (the same plots for v2 can be found here [16]). The y-axes show the top 10 most prevalent metric values, ordered from the least frequent to most frequent. Thus, the set of metric values included in the y-axis is significant (only the top ten are shown). The x-axes show the years. Each (x,y) point indicates that in year x the metric value at y has a rank indicated by the number in the circle. The size of the circle is proportional to the number of times that metric value appeared in a score in that year. For example in Figure 3, in 2017, the metric value AV-N was the fourth most frequent metric value within the set of all v3 vectors. However, in 2018 and 2019 this metric value became the third most frequent. Notice that in general, a value might appear in the top 10 of one year while not appearing in another year. Whenever that happens, we rank that particular value 11 for all the years in which it did not appear.

For v3 (see Figure 3), we observed that the same top 10 values appeared from 2016 to 2019. Furthermore, only one of those values is missing in the 2015 top 10. In addition, these values were ranked almost the same over the years. The top 2 are constant and in the same order over the time period 2015 to 2019. The top 4 and the bottom 4 (including the 11th appended value) are also constant with minor changes in the order they appear over the years. The v2 data shows similar results (see [16]). This is another illustration of the stationary threat landscape observed earlier. It also corroborates the observations in Figure 1, that the landscape has been dominated by just a few vulnerability types.

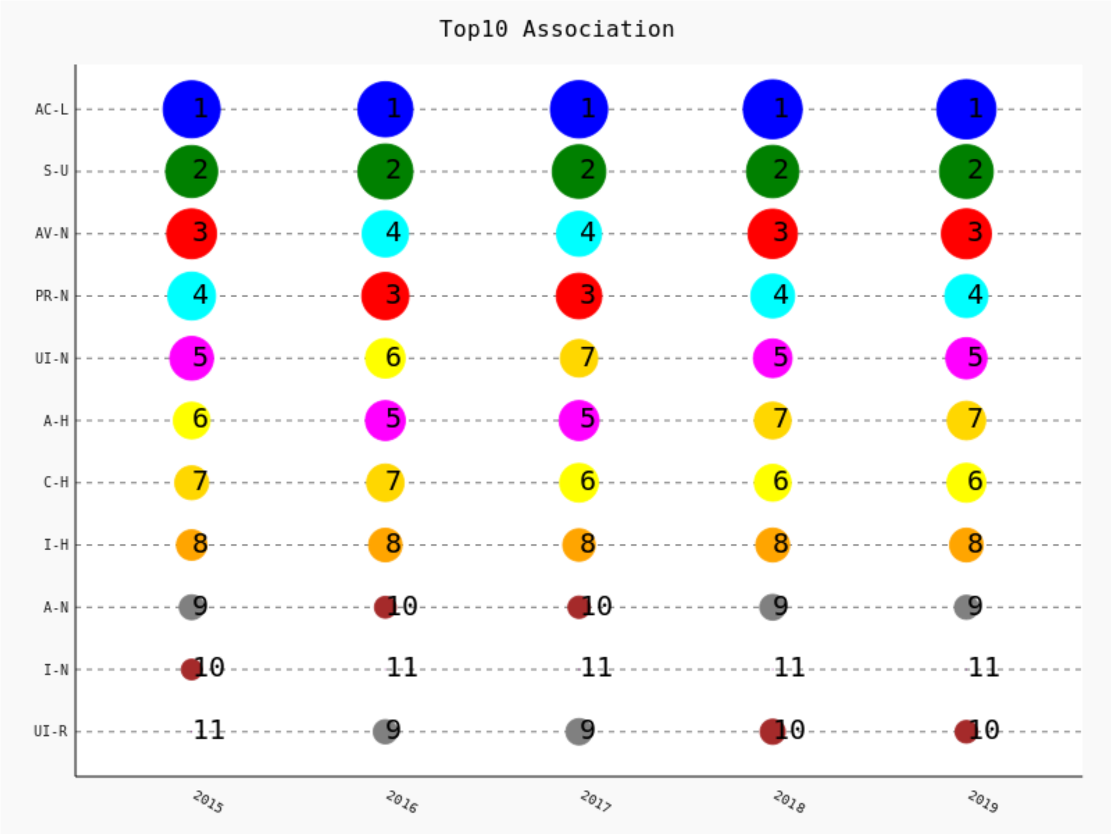


Fig. 3. CVSS v3 top 10 rankings

In conclusion, our data indicates that the vulnerability threat landscape has been dominated by a few vulnerability types and has not evolved over the years. The overwhelming majority of software vulnerabilities are exploitable over the network (i.e., remotely). The complexity to successfully exploit these vulnerabilities is dominantly low and very little authentication to the target victim is necessary for a successful attack. Moreover, most of the flaws require very limited interaction with users. The damage of these vulnerabilities has, however, mostly been confined within the scope of the compromised systems.

IV. Related Work

There are many efforts to understanding the software vulnerability landscape. These efforts include reports by security solutions vendors [18], [19], white papers from non-profits such as MITRE [20] and SANS [21], as well as academic papers. For CVSS, most studies focused on the aggregation equation that produces the CVSS numerical scores representing the severity of the vulnerability. Surprisingly, we found no studies on v3 despite its preponderance in commercial security software.

Reference [8] is among the first statistical studies of the CVSS scoring system. It evaluated v1 and proposed improvements that contributed to the release of v2. Our study considers both v2 and v3 (but does not try to improve on either). Relative to the statistical evaluation, we consider our paper as a continuation and update of the work in [8]. However, our work uses data from a much longer time period. It also goes one step further by analyzing association rules of vulnerability metrics. It is worth noting that there are similarities between the results of the two studies. For instance, both papers show the predominance of certain types of vulnerabilities. Our historical analysis (which was not performed in [8]) shows that this predominance is maintained over the years.

Reference [11] considers CVSS v1 and v2 and analyzes how effectively v2 addresses the deficiencies found in v1. It also identifies new deficiencies. In contrast, our motivation was to understand the threat landscape.

Reference [13] uses empirical data from an international cyber defense exercise to study how 18 security estimation metrics based on CVSS correlate with the actual *Time-To-Compromised* (TTC) of 34 successful attacks. This study uses TTC as a dependent variable to analyze how well different security estimation models involving CVSS are able to approximate the actual security of network systems. The results suggest that security modeling with CVSS data alone does not accurately portray the time-to-compromise of a system. This result questions the applicability of the CVSS numerical scoring equation. Our study focused on the raw CVSS vectors, which represent the actual experts' opinions about the vulnerabilities.

Reference [22] uses NVD data to study trends and patterns in software vulnerabilities in order to predict the time to next vulnerability for a given software application. Data mining techniques were used as prediction tools. The vulnerability features used to aid the prediction are the published time of each vulnerability and its version. We believe that these features are not sufficiently informative. Instead, we directly use the eight metrics from the CVSS base scores which constitute the best available information covering large multi-year sets of vulnerabilities.

Reference [23] also carried out a predictive study based on the NVD/CVSS and ExploitDB [24] data. Using the CVSS data, it attempts to answer two questions: *(1) Can we predict the time until a proof of concept exploit is developed based on the CVSS metrics? and (2) Are CVSS metrics populated in time to be used meaningfully for exploit delay prediction of CVEs?* The former is answered in the positive, while the latter is answered in the negative. While using the same datasets, our objective differs from that in [23]. We did not attempt to predict the threat landscape; we provide a thorough historical and statistical study of vulnerabilities for the last fifteen years.

The work in [25] is another assessment of CVSS. It evaluates the trustworthiness of CVSS by considering data found in five vulnerability databases: NVD, X-Force, OSVDB (Open Source Vulnerability Database), CERT-VN (Computer Emergency Response Team, Vulnerability Notes Database), and Cisco IntelliShield Alerts. It then uses a Bayesian model to study consistencies and differences. It concluded that CVSS is trustworthy and robust in the sense that most of the databases generally agree. This suggests that our focus on the NVD to study the threat landscape is justified: studies using data from the other databases will likely lead to the same conclusions.

All of the studies cited above are focused on v1 and v2. In our literature survey, we did not find a single study that uses the updated and significantly modified v3 to understand the software vulnerability landscape. We believe that the present paper is the first of this kind in doing so. Furthermore, our study is the first to use association rule mining and co-occurrence of vulnerability metrics' values in an attempt to characterize the software threat landscape.

V. Conclusion

Our data indicates that the vulnerability threat landscape for publicly disclosed vulnerabilities has been dominated by a few vulnerability types and has not significantly changed from 2005 to 2019. However, the underlying software flaw types that enable these vulnerabilities change dramatically from year to year (for example, see [26]). This means that many flaw types result in vulnerabilities with the same properties. This is bad news because it means, as a security community, it will be difficult to eliminate certain vulnerability types because they result from a plethora of underlying software flaw types.

Another concern is that the overwhelming majority of software vulnerabilities are exploitable over the network. When developing software, efforts should be made to reduce unnecessary connections, protect necessary ones, and require more authentication where possible to reduce attack surface area. Another significant issue is that most of the vulnerabilities require no sophistication to be exploited (but again this is hard to improve upon due to the many software flaw types that allow this).

These two factors together combined with the finding that most vulnerabilities require very limited interaction with users facilitates the widespread hacking occurring today. Often in security literature the human is cited as the weakest link. While certainly humans can be exploited, within the set of CVE type vulnerabilities, exploitation of humans plays a very minor role. Hence, although training humans might always help strengthen security, to obtain a better impact in this area, the priority should be shifted to correcting these constant vulnerabilities.

Overall, this study documents the security community's inability to eliminate any types of vulnerabilities through addressing the related software flaw types. In 15 years, the vulnerability landscape has not changed; through the lens of the metrics in this paper we are not making progress. Perhaps we as community need to "stop and think" about the ways we are developing software and/or the methods we use to identify vulnerabilities. The security community needs new approaches. We do not want to write this same paper 15 years from now showing that, once again, nothing has changed.

Overall, this study shows that either we (the community) are incapable of correcting the most common software flaws, or we are focusing on the wrong flaws. In either case, it seems to us that there is a need to "stop and think" about the ways we are developing software and/or the methods we use to identify vulnerabilities.

Acknowledgement

This work was partially accomplished under NIST (National Institute for Standards and Technology) Cooperative Agreement No.70NANB19H063 with Prometheus Computing, LLC. The authors would like to thank the NVD staff for their review and consideration of this work.

Any mention of commercial products or entities is for information only; it does not imply recommendation or endorsement.

References

[1] "National vulnerability database," 2020, accessed: 2020-01-10. [Online]. Available: https://https//nvd.nist.gov

[2] "Common vulnerability scoring system special interest group," accessed: 2019-12-10. [Online]. Available: https://www.first.org/cvss

[3] "Forum of incident response and security teams," accessed: 2020-01-10. [Online]. Available: https://www.first.org/

[4] M. Schiffman, A. Wright, D. Ahmad, and G. Eschelbeck, "The common vulnerability scoring system," *National Infrastructure Advisory Council, Vulnerability Disclosure Working Group, Vulnerability Scoring Subgroup*, 2004.

[5] P. Mell, K. Scarfone, and S. Romanosky, “A complete guide to the common vulnerability scoring system version 2.0,” in *Published by FIRST-Forum of Incident Response and Security Teams*, vol. 1, 2007, p. 23.

[6] “Common vulnerability scoring system v3.0: Specification document,” accessed: 2020-2-5. [Online]. Available: https://www.first.org/cvss/v3.0/specification-document

[7] “Common vulnerability scoring system v3.1: Specification document,” accessed: 2020-2-5. [Online]. Available: https://www.first.org/cvss/v3.1/specification-document

[8] P. Mell and K. Scarfone, “Improving the common vulnerability scoring system,” *IET Information Security*, vol. 1, no. 3, pp. 119–127, 2007.

[9] P. Mell, K. Scarfone, and S. Romanosky, “Common vulnerability scoring system,” *IEEE Security & Privacy*, vol. 4, no. 6, pp. 85–89, 2006.

[10] H. Holm and K. K. Afridi, “An expert-based investigation of the common vulnerability scoring system,” *Computers & Security*, vol. 53, pp. 18–30, 2015.

[11] K. Scarfone and P. Mell, “An analysis of cvss version 2 vulnerability scoring,” in *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*. IEEE Computer Society, 2009, pp. 516–525.

[12] R. Wang, L. Gao, Q. Sun, and D. Sun, “An improved cvss-based vulnerability scoring mechanism,” in *2011 Third International Conference on Multimedia Information Networking and Security*. IEEE, 2011, pp. 352–355.

[13] H. Holm, M. Ekstedt, and D. Andersson, “Empirical analysis of system-level vulnerability metrics through actual attacks,” *IEEE Transactions on dependable and secure computing*, vol. 9, no. 6, pp. 825–837, 2012.

[14] 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.

[15] “Common vulnerabilities and exposures,” 2020, accessed: 2020-2-5. [Online]. Available: https://cve.mitre.org

[16] A. Gueye and P. Mell, “A historical and statistical studyof the software vulnerability landscape,” 2021.

[17] NVD, “private communication,” Mar. 2019.

[18] Symantec, “2019 internet security threat report,” 2020, accessed: 2020-02-01. [Online]. Available: https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf

[19] McAfee, “Mcafee labs 2019 threats predictions report,” 2020, accessed: 2020-02-01. [Online]. Available: https://www.mcafee.com/blogs/other-blogs/mcafee-labs/mcafee-labs-2019-threats-predictions/

[20] MITRE, “2019 cwe top 25 most dangerous software errors,” 2020, accessed: 2020-02-01. [Online]. Available: https://cwe.mitre.org/top25/archive/2019/2019_cwe_top25.html

[21] SANS, “2020 sans cyber threat intelligence (cti) survey,” 2020, accessed: 2020-02-01. [Online]. Available: https://www.sans.org/reading-room/whitepapers/analyst/2020-cyber-threat-intelligence-cti-survey-39395

[22] S. Zhang, D. Caragea, and X. Ou, “An empirical study on using the national vulnerability database to predict software vulnerabilities,” in *International Conference on Database and Expert Systems Applications*. Springer, 2011, pp. 217–231.

[23] Y. Y. A. Feutrill, D. Ranathunga and M. Roughan, “The effect of common vulnerability scoring system metrics on vulnerability exploit delay,” in *2018 Sixth International Symposium on Computing and Networking (CANDAR), Takayama*, 2018, pp. 1–10.

[24] “Exploit database,” 2020, accessed: 2020-02-01. [Online]. Available: https://www.exploit-db.com/

[25] P. Johnson, R. Lagerstrom, M. Ekstedt, and U. Franke, “Can the common vulnerability scoring system be trusted? a bayesian analysis,” *IEEE Transactions on Dependable and Secure Computing*, 2016.

[26] “National vulnerability database, cwe over time,” 2019, accessed: 2019-12-10. [Online]. Available: https://nvd.nist.gov/general/visualizations/vulnerability-visualizations/cwe-over-time

Seismic and Durability Assessment of Externally Bonded FRP Retrofits in Reinforced Concrete Structures after 2018 Anchorage, AK Earthquake

Sandra Milev[1], Shafique Ahmed[2], Mariam Hassan[3], Siamak Sattar[3], David Goodwin[3] and Jovan Tatar[1]

[1] University of Delaware, Newark DE, USA
[2] Echem Consultants LLC, Poughkeepsie, NY, USA
[3] National Institute of Standards and Technology, Gaithersburg, MD, USA

jtatar@udel.edu

Abstract. Externally bonded fiber-reinforced polymer (EBFRP) composites are a cost-effective material used for repair and seismic retrofit of existing concrete structures. Even though EBFRP composites have been extensively utilized over the past 20 years as seismic retrofits, there are few data documenting their performance in a real shaking event or after long-term use on concrete structures. In this study, semi-destructive and non-destructive techniques were employed to evaluate the performance and durability of EBFRP-retrofitted buildings that had experienced the 2018 Cook Inlet Earthquake in Anchorage, AK. The performance of EBFRP was evaluated and documented through photographic evidence. Acoustic sounding, infrared thermography, and bond pull-off tests were utilized to evaluate the quality of bonding between the EBFRP and concrete. EBFRP samples were also collected from building interiors and exteriors for chemical and thermal analysis to evaluate the long-term effects of environmental exposure. Although environmental conditions were found to influence the bond quality between the EBFRP composite and concrete substrate, no major signs of earthquake damage to the building components retrofitted with EBFRP were noted. Materials characterization results demonstrated no evidence of polymer matrix degradation in exterior EBFRP samples.

Keywords: EBFRP retrofit; seismic; deterioration; reconnaissance; durability; materials characterization.

1 Introduction

Externally bonded fiber-reinforced polymer (EBFRP) composites provide effective solutions for rehabilitating and strengthening existing concrete structures that have suffered deterioration or require seismic retrofit. EBFRP composites possess several properties that make them an attractive retrofitting solution, including their lightweight, high strength, ability to conform to existing structural components geometries, and corrosion resistance. Over the past 30 years, there have been numerous pub-

Milev, Sandra; Goodwin, David; Sattar, Siamak; Ahmed, Shafique; Tatar, Jovan. ”Performance of Externally Bonded Fiber-Reinforced Polymer Retrofits in Reinforced Concrete Structures during the 2018 Anchorage, Alaska Earthquake.” Presented at 10th International Conference on FRP Composites in Civil Engineering (CICE 2020/2021), Istanbul, TR. July 01, 2020 - July 03, 2020.

lications to evaluate the effectiveness and benefits of EBFRP for the strengthening of reinforced concrete structural members. These studies showed that application of EBFRP composites can provide performance improvements including enhanced confinement and ductility, load-bearing capacity, blast resistance, and shear resistance (Bakis et al. 2002; Buchan and Chen 2007; He et al. 2015; Ma et al. 2016). However, there are not many data documenting performance of EBFRP in a real-world earthquake. Furthermore, the long-term performance of EBFRP is an issue that remains a concern in the engineering community (Goodwin et al. 2019). Evidence regarding the long-term performance of EBFRPs is currently based on limited short-term accelerated conditioning testing (Tatar and Hamilton 2014) without validation from real-world outdoor testing. Due to the effect of combined environmental conditions and factors, field exposure can result in different rates and mechanisms of degradation than those determined on the basis of controlled laboratory experiments. In addition, design recommendations, informed by laboratory experiments only, can result in overly-conservative designs that can negatively impact the cost-effectiveness of a retrofitting system (Zhang et al. 2014). Reports that describe the long-term performance of EBFRP composites in the field provide valuable information but are often limited by their use of a single type of material, a single retrofit configuration, a single outdoor exposure site, and the duration of the study. Some early projects and demonstrations have shown that EBFRP composites can have excellent durability characteristics in warm and cold climates for several years (Sheikh and Tam 2007; Steckel and Hawkins 2005). A field study conducted on carbon fiber reinforced polymer (CFRP)-wrapped girders that were taken out of service from bridges located in Florida indicated that bonded CFRP repairs can last upwards of 15 years and perhaps beyond with an EBFRP system applied to a well-prepared substrate surface (e.g., damaged concrete repaired if needed and surface treated to achieve the recommended ICRI surface profile (ICRI-No.03732 1997), using proper installation techniques that include full saturation, environmental conditions that do not affect epoxy cure, and removal of air voids by hand or with a trowel (Hamilton et al. 2017; Tatar et al. 2016). To contribute to existing field studies, this research project was initiated with an objective of evaluating field performance of EBFRP under earthquake loading and the effects of Alaska's subarctic climate on EBFRP durability.

2 Studied buildings

Past seismic activity in Anchorage, AK led the community to retrofit many buildings over the years – some of them with EBFRP. During the 2018 Cook Inlet Earthquake, several buildings in Anchorage, AK retrofitted with EBFRP experienced shaking. For this reason, these buildings were visually inspected and two of them were selected for more detailed evaluation that is presented in this paper – the McKinley Tower (MKT) and Ted Stevens International Airport (TSIA). The epicenter of the earthquake that occurred on November 30th 2018 was 12 km north of Anchorage, AK. Based on the response spectra from several locations in Anchorage, AK, the 7.1 M_w earthquake had an intensity lower than the design earthquake. Most notable damage was caused by geotechnical failures that affected highways and many roads in the area. Reports of post-earthquake inspection suggest that damage to the buildings was mostly limited to non-structural elements (StEER and EERI, 2018).

MKT is a 36.6 m (120 ft) tall reinforced concrete residential building in Anchorage, constructed in 1951. The building was severely damaged in the 1964 Great Alaskan earthquake (M 9.2). After the earthquake, the building was abandoned and left unoccupied for approximately twenty years. It had fallen behind seismic codes and needed a retrofit before it could be reoccupied. A traditional retrofitting project was started in 1998 and then abandoned shortly thereafter because of its high cost. A new seismic evaluation was performed in 2004 when EBFRP was selected as a cost-effective solution to retrofit and strengthen the structure. Carbon and glass EBFRP retrofits, hereby referred to as CFRP and GFRP retrofits, respectively, were applied to columns, walls, and beams. The majority of the EBFRP retrofits were applied on floors 5 through 14 (Ehsani 2017).

Terminal 2 of TSIA was built in 1968 or later and seismically retrofitted in July 2008. The existing exterior columns in Terminal 2 were retrofitted to act as the boundary elements for the new shear walls. An interior column in a baggage handling area was also retrofitted. For both interior and exterior applications, CFRP was used to confine the columns.

During their lifetime, some of the outdoor EBFRP retrofits at MKT and TSIA were exposed to harsh environmental conditions typical for Anchorage, AK – average daytime summer temperatures from 12 ℃ to 25 ℃, and average daytime winter temperatures in the -15 ℃ to 0 ℃ range, with an average relative humidity of 70 % (National Oceanic and Atmospheric Administration 2020).

3 Experimental program

3.1 Phase 1: Field assessment

Prior to field assessment, data and information was collected about the buildings retrofitted with EBFRP in Anchorage, AK. The field assessment then took place in January 2019. All building inspections involved photographic evidence collection and further gathering of information related to EBFRP design documentation, details about EBFRP materials, and inspection reports from before and after the repair. Since the effectiveness of EBFRP strengthening depends on the properties of the bond between the concrete and EBFRP, a detailed evaluation of the EBFRP bond was conducted using acoustic sounding, infrared (IR) thermography, and bond pull-off tests at Ted Stevens International Airport and McKinley Tower.

Acoustic sounding and *Infrared (IR) thermography* were used to locate areas of EBFRP debonding from the concrete substrate (Reay and Pantelides 2006). Debonded areas detected by acoustic sounding, in this case—light tapping with a hammer—have a hollow sound upon hammer impact. After detection, hollow-sounding areas were marked with a blue tape and documented with photographs. For IR thermography, the surface of the EBFRP was evenly heated by a halogen lamp and the heat dissipation was measured using an IR thermography instrument from an approximately consistent distance. The technique works by detecting the slower heat dissipation through the debonded areas than through the areas that are properly bonded. The debonded areas show up as "hot spots" in an IR image. Although IR thermography is a useful tool to detect debonding, the technique is time-consuming (*e.g.*, an hour or more per column face). For this reason, the IR thermography method was used only for one interior and one exterior column at TSIA. During post-processing of the IR thermography data, the size and shapes of the debonded areas were matched with the debonded areas identified with acoustic sounding. Acoustic sounding data was also used to identify "false positive" readings in the thermographs associated with the uneven thickness of the epoxy adhesive or EBFRP composite (Brown, 2005).

Bond pull-off tests were conducted according to the ASTM D7522 (2015). A total of 12 pull-off tests were conducted in bonded areas identified with acoustic sounding. One pull-off test per FRP row was conducted at locations away from edges, to avoid areas that may have contained higher stress concentrations. Although EBFRPs retrofitted to columns are considered contact-critical and do not require a strong bond between the EBFRP and the concrete substrate as is needed in bond critical applications, the EBFRP material used to retrofit the columns was the same material used in bond critical applications and only bonded areas of the EBFRP retrofitted column were evaluated. In short, the test procedure consisted of: (1) sanding the EBFRP surface to remove any paint or coatings followed by cleaning with acetone to eliminate surface contamination; (2) adhering 5-cm aluminum pucks to the prepared test area using a quick-set epoxy adhesive; (3) coring the EBFRP around the perimeter of each puck with a 5-cm core-saw; (4) clamping the adhesion tester to the puck; and (5) applying a tensile load at a constant loading rate of approximately 335 N/s. The test setup is shown in Figure 1 and the test procedure and failure modes evaluated were in accordance with ASTM D7522 (2015).

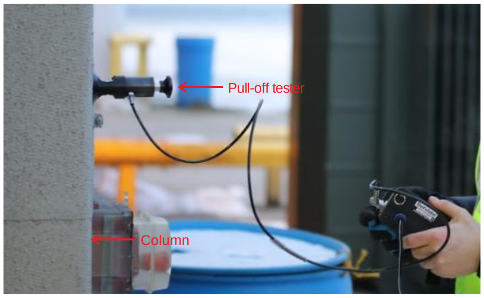


Figure 1. Pull-off test setup

3.2 Phase 2: Materials characterization

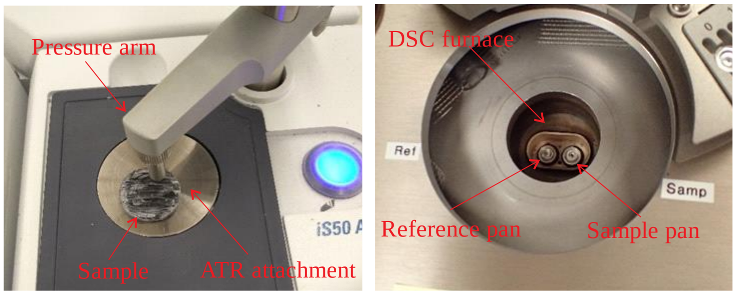
Differential scanning calorimetry (DSC) and *Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy* (ATR-FTIR) data were collected from CFRP and GFRP samples obtained from MKT and TSIA. Measurements were conducted at the Center for Composite Materials at the University of Delaware and the National Institute of Standards and Technology (NIST).

DSC was used to measure the glass transition temperature (T_g) values of the EBFRP samples to evaluate the effect surrounding temperature could have on the mechanical properties of the retrofit as well as the degree of polymer cross-linking density. The glass transition is the temperature range over which a polymer transitions from a solid to rubbery state, which is accompanied by decreased strength and stiffness. In DSC analysis, the EBFRP sample and a reference material (Figure 2) are heated at the same rate in a temperature-controlled furnace, and the heat flow difference between the sample and reference is recorded. A step in the heat flow curve indicates the glass transition temperature range.

For DSC experiments, triplicate circular specimens with a mass between 8 mg and 16 mg were punched out from individual EBFRP samples using a press, and then placed into aluminum pans. To ensure seamless heat transfer between the pan and the material inside, mineral oil was used. The samples were covered with a pierced aluminum lid and sealed with a manual press. Each DSC experiment involved heating the samples from -20 ℃ to 250 ℃ at a constant heating rate of 10 ℃/min under a nitrogen atmosphere. The T_g was determined as the midpoint temperature in the first heat run according to ASTM E1356 (2008).

ATR-FTIR spectroscopy was used to determine if any chemical degradation of the EBFRP composite retrofits had occurred over long-term outdoor exposure. ATR-FTIR spectroscopy is based on the interaction between infrared light and a sample. This technique relies on molecular vibrations in a sample to identify the presence of certain functional groups. A functional group absorbs infrared light if frequencies of the light and molecular vibrations are equal. As a result, a spectrum that shows absorbance as a function of wavelength is generated.

The spectra were collected using an FTIR spectrometer (Nicolet iS50 FTIR) with a diamond crystal ATR accessory. Each spectrum collected represents an average of 256 scans with a resolution of 4 cm^{-1}. ATR-FTIR spectra were baseline-corrected in Origin software. All spectra were normalized to the 1508 cm^{-1} band of the interior MKT sample corresponding to benzene ring stretching that was present across all spectra and is not expected to change with environmental exposure. At least three replicate areas per sample were measured and the replicate spectra were averaged after baseline correction. The test setup is shown in Figure 2.

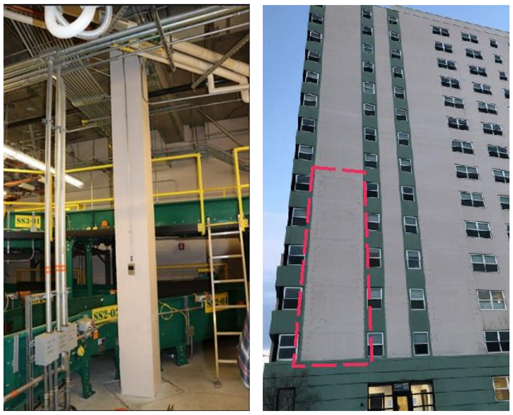


a) b)

Figure 2: a) ATR-FTIR test setup, and b) sample inside DSC chamber

4 Results and Discussion

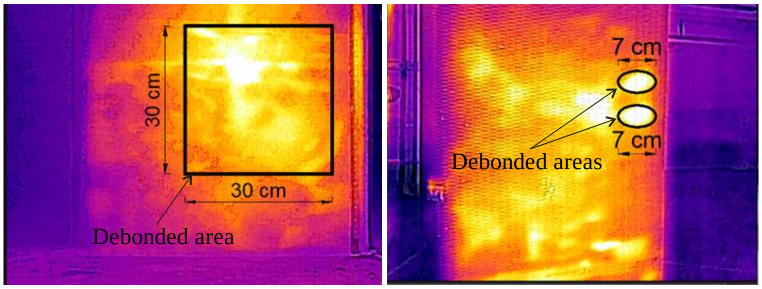
Field assessment data acquired through visual inspection of eight representative buildings retrofitted with EBFRP showed no visible signs of earthquake damage to EBFRP composite retrofits. Figure 3a shows an example of an interior column at TSIA. Interior columns, wrapped with 2 plies of CFRP for confinement, were undamaged by the earthquake. For MKT, an exterior shear wall retrofitted with GFRP is shown in Figure 3b. Wall boundary elements were formed by applying horizontally oriented unidirectional GFRP and CFRP on three sides of the window corner opening. Additional bolts were installed through the wall to confine the boundary elements. No obvious earthquake damage to the EBFRP retrofits was observed in the post-earthquake inspections. It should be noted that the team inspected only visible, exposed, EBFRP retrofits; during inspections of many of the buildings, the team faced accessibility problems as many EBFRP retrofits were located behind drywall, panels, or other architectural finishes.



a) b)

Figure 3. Examples of typical EBFRP retrofits: a) interior column at TSIA; b) shear wall at MKT retrofitted with GFRP (marked by a dashed red line).

Acoustic sounding and *Infrared (IR) thermography*. Several debonded areas of varying size and shape were identified within the EBFRP composite system using acoustic sounding and the debonded areas were marked with tape. Qualitative IR thermography measurements were then taken and the debonded areas identified with tape were later matched up to IR thermography images (Figure 4). On the interior column, thirteen relatively small debonded areas (between 10 cm^2 to 40 cm^2) were detected. In contrast, four larger debonded areas that exceeded 160 cm^2 were observed on the exterior column. According to ACI 440.2R (2017), debonded areas larger than 160 cm^2 can affect the performance of EBFRP retrofits and should be repaired by selectively cutting away the affected sheet and applying an overlapping sheet patch of equivalent plies. It is unknown whether these debonded areas formed because of environmental deterioration, or due to construction defects. But, considering that debonding was more severe on the exterior column compared to the interior column and presuming that surface preparation and construction methods were the same between the columns, we conclude that environmental conditions likely had some influence on the EBFRP debonding in the exterior column. For field investigations, initial bond durability data would be useful as a baseline for researchers assessing bond quality after long-term outdoor exposure and hazard events.



a) b)

Figure 4. Thermal IR image of debonding at TSIA: a) an exterior column and b) interior column. Bright areas were marked based on the location of tape from acoustic sounding.

Bond pull-off tests. Bond pull-off strengths were compared against the ACI 440.2R minimum bond strength requirement of 1.4 MPa (Figure 5). Initial bond pull-off strength data was not available for any of the building components investigated. In future FRP retrofit installations, it would be useful to obtain and maintain this baseline data for long-term field inspections. Four different failure modes were observed in the bond pull-off tests (ASTM D7522): cohesive failure in the concrete substrate; adhesive failure at the EBFRP/concrete interface; mixed adhesive and cohesive failure; and, bonding adhesive failure at the loading fixture. Representative photographs of different failure modes are shown in Figure 6. Cohesive failure in concrete (Figure 6a) is desirable as it indicates sound adhesion between EBFRP and concrete, while the failure modes in Figure 6b and Figure 6c indicate improper adhesion between EBFRP and concrete, which can be caused by improper concrete surface preparation, environmental conditioning, or a combination of the two. Bonding adhesive failure at the loading fixture (Figure 6d) occurred when the epoxy used to attach the fixture did not fully cure by applied heat in the extremely cold temperatures (< -12 ℃) experienced during inspection. This failure mode was a non-result and was mostly avoided during subsequent pull-off tests on the same building component by applying heat for a longer period of time. Pull-off test strength data for this failure mode was not useful and is thus not included in Figure 5.

At TSIA, measured bond pull-off strengths, both on the exterior and interior columns, were lower than the minimum requirement. Of note, pull-off bond strengths measured on the exterior column were lower compared to the interior column. Furthermore, two adhesive failures at the EBFRP-adhesive interface, which is an indication of poor bond properties, were recorded for the exterior column. The lower bond pull-off strengths on the exterior column at TSIA and the presence of adhesive failure modes provide evidence that there are some issues in regard to the durability of the bond between the CFRP and concrete.

All pull-off tests at MKT passed the 1.4 MPa minimum bond strength recommendation. Cohesive failure modes were observed in all experiments (with the exception of one bonding adhesive failure of the fixture on the exterior of the building) indicating good adhesion properties between the GFRP and concrete. Better bond performance observed at MKT in comparison to TSIA could be due to the mismatch in coefficients of thermal expansion (CTE) between the concrete and CFRP at TSIA. Carbon fibers have a negative CTE which leads to expansion under low ambient temperatures and, consequently, differential movement between the concrete and the EBFRP. On the other hand, the CTE of glass fibers is similar to the CTE of concrete, resulting in compatible deformations.

Several limitations should be noted for bond degradation assessment. First, with pull-off tests, the sample size was small due to the number of loading fixtures available at the time of this study and the cold weather issues that led to some bonding adhesive failures at the loading fixture. Further investigation of pull-off tests using a larger sample size is underway using data from a return trip to Anchorage. In future studies, improved statistical analysis can be accomplished using at least three pull-off tests per building component, with more pull-off tests conducted if inconsistencies are observed. Investigation of more than one building component should also be considered when it is possible in the field, and this approach was taken on our return trip to Anchorage. In comparison to the IR thermography technique, pull-off tests had the disadvantage of being destructive and sometimes led to inconsistent results. However, IR thermography had a longer measurement time and yielded false positives from thickened adhesive regions. Nevertheless, IR thermography could be validated with acoustic sounding and less localized results could be obtained with IR thermography compared to the highly localized results obtained with pull-off tests. Overall, a balanced approach with multiple techniques can help validate the findings of each individual technique.

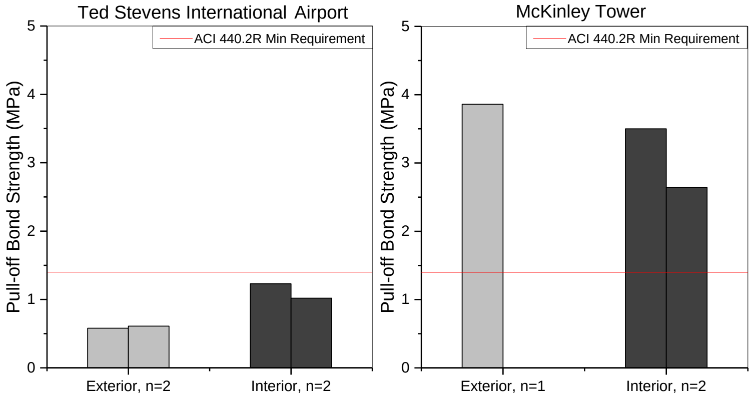


Figure 5. Bond pull-off test strengths for all failure modes except bonding adhesive failure at the loading fixture. Each bar represents one pull-off test strength and the red line indicates the minimum bond pull-off test strength according to ACI 440.2R. Only one replicate was possible for the MKT exterior at the time of this field study because of a bonding adhesive failure at the loading fixture for the first pull-off test conducted.

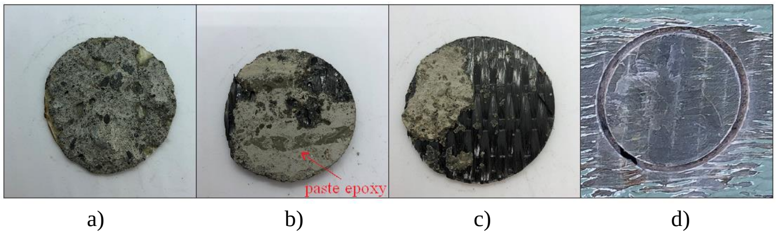


Figure 6. Typical bond pull-off test failure modes: a) cohesive failure in the concrete substrate, b) adhesive failure at the EBFRP/concrete interface, c) mixed adhesive and cohesive failure mode, and d) bonding adhesive failure at the fixture-EBFRP interface.

Thermal analysis–The average T_g values obtained for CFRP retrofits from the interior and exterior column at TSIA are shown in Figure 7. The observed T_g values for exterior CFRP retrofits at TSIA were in the 56 ℃ to 58 ℃ range. Interior retrofits at TSIA had lower T_g values which varied from 47 ℃ to 56 ℃. The T_g values were likely greater for the exterior samples since higher outside temperatures during the summer season may have stimulated post-curing of the epoxy, leading to more cross-linking density.

The T_g values of GFRP samples from MKT were in the 53 ℃ to 54 ℃ range with no statistically significant difference between interior and exterior samples. During the service life of GFRP samples from MKT, the highest interior and exterior ambient temperatures most likely reached similar values as a result of less strictly controlled ambient conditions typical for residential buildings. Taking into account this assumption, similar values of measured T_g on the interior and exterior of MKT, can be attributed to a similar degree of post-curing.

According to ACI 440.2R, for a dry environment, it is suggested that the anticipated maximum service temperature of an EBFRP system not exceed (T_g – 15 ℃). ACI 440.2R guidelines do not recommend specific service temperatures for different climatic areas, leaving it up to the licensed designed professional to specify the maximum service temperature. In this study, the maximum interior service temperature was conservatively assumed to be 30 ℃ for TSIA and MKT. The maximum service temperature for exterior applications was assumed to be 55 ℃ based on the literature (Michels et al. 2015). In the referenced study, the measurements performed on a pedestrian bridge in Switzerland, directly exposed to direct sunlight, show that the surface temperature was 55 ℃ which was significantly greater than the measured ambient temperature on the day of the measurement (33 ℃).

The minimum recommended exterior T_g (55 ℃+ 15 ℃), based on the previously described exterior service temperature and the ACI 440.2R recommendation, was compared to the measured T_g values at MKT and TSIA. The measured T_g values of the exterior samples did not satisfy the minimum criteria from ACI 440.2R, for the suggested exterior service temperature of 55 ℃, which may have adversely affected the mechanical properties of the EBFRP retrofits. In contrast, the measured interior T_g values were greater than the recommended minimum interior T_g (30 ℃+ 15 ℃) which indicates that there was no impact of T_g on the mechanical properties of the interior EBFRP retrofits at both MKT and TSIA.

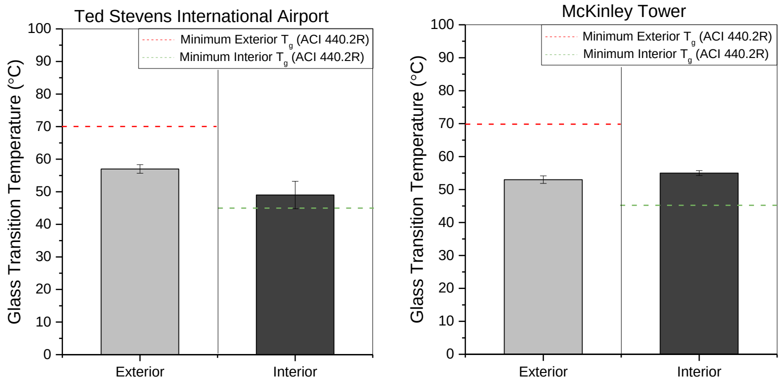
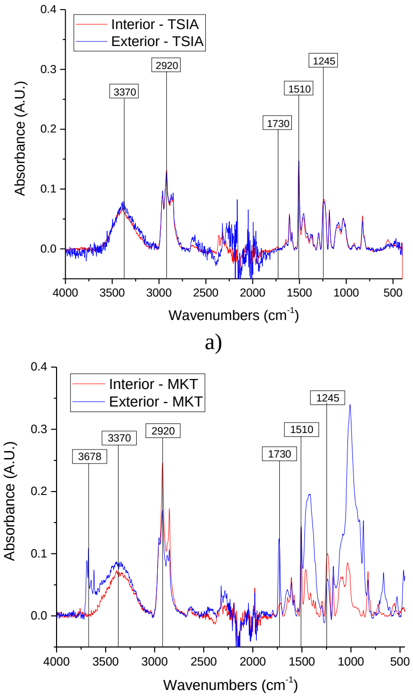


Figure 7. The average T_g values of triplicate (8 mg to 16 mg) circular punched-out EBFRP specimens from individual samples collected at MKT and TSIA (error bars indicate one standard deviation). The minimum interior and exterior values indicate the threshold at which the T_g value of the EBFRP must exceed to maintain optimal mechanical properties.

ATR-FTIR. Typical spectra collected on the exterior and interior EBFRP samples are shown in Figure *8*. The peaks at 2920 cm^{-1} and 2850 cm^{-1} are related to C-H bonds of the monomer units, the peak at 1245 cm^{-1} corresponds to C-O bonds in epoxy, the peak at 1510 cm^{-1} corresponds to benzene ring, and the broad peak at 3370 cm^{-1} is related to hydroxyl groups (Cysne Barbosa et al. 2017). The band at 1730 cm^{-1}, when present in the spectra, corresponds to the carbonyl group and can indicate oxidation (Rezig et al. 2006).

The results suggest that the epoxy matrix at TSIA did not chemically change due to the field exposure to the subarctic climate: the spectra are qualitatively similar between the interior and exterior sample. Further analysis is underway with newly collected CFRP samples from TSIA to determine if any chemical changes are present among several other retrofitted columns. For the exterior EBFRP sample from MKT, the glass fibers were poorly saturated with epoxy, making it difficult to investigate epoxy degradation. There were some dissimilarities in peaks and peak intensities between the interior and exterior GFRP samples. The exterior GFRP sample was thought to be contaminated—the peaks around 3680 cm^{-1} likely correspond to silica sand particles in the paste adhesive that was used to bond the EBFRP to the concrete substrate. The carbonyl peak at 1730 cm^{-1} was also present in the samples; the research team is currently conducting additional experiments to determine whether this peak is associated with oxidation of the epoxy matrix or is caused by contaminants that may have been present in the specimen (*e.g.*, paint). Further analysis is underway with newly collected GFRP samples from MKT exterior.



a)

b)

Figure 8. Qualitative comparison of spectra on exterior and interior EBFRP from: a) TSIA, and b) MKT. Each spectrum is the average of three spectra from different areas of the same EBFRP sample. Bands of interest are denoted in boxes with units of cm^{-1}.

5 Summary and Conclusions

The primary objective of this work was to provide information on the performance of EBFRP retrofits in the 2018 earthquake in Anchorage, AK and the durability of

EBFRP retrofits after long-term exposure to Alaska's sub-arctic climate. EBFRP retrofits at two concrete structures, the McKinley Tower and Ted Stevens International Airport, retrofitted in 2004 and 2008, respectively, were investigated. Field assessment and laboratory testing of CFRP and GFRP samples collected from the buildings were performed. Field assessment included visual inspection to evaluate visible post-earthquake damage, IR thermography and acoustic sounding to detect debonded areas, and bond pull-off tests to evaluate the EBFRP-concrete bond quality. EBFRP material samples were collected during site visits to conduct DSC measurements to determine the glass transition temperature of the polymer matrix, and ATR-FTIR measurements to identify possible chemical degradation in the composite samples. Based on the experimental results presented in this paper, the following conclusions were made:

- No apparent signs of earthquake damage to EBFRP-retrofitted components were observed. Considering that many structural elements were located behind drywalls, visual inspection was not always possible. IR thermography and bond pull-off tests indicated that bond degradation may have occurred on exterior retrofits. It is not possible to draw firm conclusions about bond degradation without baseline data regarding the original quality of the bond and with the small sample size of pull-off tests. On a return trip to Anchorage, a larger sample size of three bond pull-off tests per column for multiple columns was accomplished, with more tests conducted when inconsistencies were observed. Further analysis of this data is underway.
- IR thermography required validation of debonded areas with acoustic sounding and was a time-consuming process. However, it was not destructive like pull-off testing and was more representative of the total column area. Overall, the use of multiple techniques can help validate the results of each individual technique and account for its limitations.
- Interior T_g values passed the minimum recommended T_g value per ACI 440.2R. However, T_g values measured on exterior EBFRP retrofits were below the minimum recommended temperatures.
- ATR-FTIR analysis shows that the polymer matrix in TSIA samples did not qualitatively changed during long-term exposure to environmental conditions in Alaska. GFRP samples from MKT showed the possible presence of contaminants in the spectra. Further investigation is underway with newly collected GFRP samples to verify the findings.

6 Acknowledgements

The authors gratefully acknowledge the funding provided by the United States National Science Foundation under the award number 1916972.

***Disclaimer*:** Certain commercial products or equipment described in this paper are present in order to adequately specify the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Institute of

Standards and Technology, nor does it imply that it is necessarily the best available for the purpose.

References

1. ACI 440.2R. (2017). “Guide for the Design and Construction of Externally Bonded FRP Systems for Strengthening Concrete Structures.” American Concrete Institute, Farmington Hills, MI 48331
2. ASTM D7522. (2015). “Standard test method for pull-off strength for FRP bonded to concrete substrate.” ASTM International, West Conshohocken, PA.
3. ASTM E1356. (2008). “Standard Test Method for Assignment of the Glass Transition Temperatures by Differential Scanning Calorimetry.” ASTM International, West Conshohocken, PA.
4. Bakis, C. E., Bank, L. C., Brown, V. L., Cosenza, E., Davalos, J. F., Lesko, J. J., Machida, A., Rizkalla, S. H., and Triantafillou, T. C. (2002). “Fiber-reinforced polymer composites for construction - State-of-the-art review.” Journal of Composites for Construction, 6(2), 73–87.
5. Brown, J. R. (2005). “Infrared Thermography Inspection of Fiber-Reinforced Polymer Composites Bonded to Concrete.” PhD Dissertation, University of Florida, Gainesville, FL, USA.
6. Buchan, P. A., and Chen, J. F. (2007). “Blast resistance of FRP composites and polymer strengthened concrete and masonry structures - A state-of-the-art review.” Composites Part B: Engineering, 38(5–6), 509–522.
7. Cysne Barbosa, A. P., Fulco, A. P., Guerra, E., Arakaki, F., Tosatto, M., Maria, M. C., and José, J. D. (2017). “Accelerated aging effects on carbon fiber/epoxy composites.” Composites Part B: Engineering, 110, 298–306.
8. Ehsani, M. (2017). “Fiber Reinforced Polymers: Seismic Retrofit of the McKinley Tower.” Structure Magazine.
9. Goodwin, D. G., Sattar, S., Dukes, J. D., Kim, J. H., Sung, L. P., and Ferraris, C. C. (2019). “Research Needs Concerning the Performance of Fiber Reinforced (FR) Composites Retrofit Systems for Buildings and Infrastructure.” Special Publication (NIST SP) - 1244.
10. Hamilton, H. R., Brown, J., Tatar, J., Lisek, M., and Brenkus, N. R. (2017). “Durability Evaluation o f Florida’s Fiber-Reinforced Polymer (FRP) Composite Reinforcement for Concrete Structures.” Florida Department of Transportation.
11. He, R., Yang, Y., and Sneed, L. H. (2015). “Seismic Repair of Reinforced Concrete Bridge Columns: Review of Research Findings.” Journal of Bridge Engineering, 20(12), 04015015.
12. ICRI-No.03732. (1997). “Selecting and Specifying Concrete Surface Preparation for Sealers, Coatings, and Polymer Overlays.” International Concrete Repair Institute, Des Plaines, IL, USA.
13. Ma, C. K., Apandi, N. M., Yung, S. C. S., Hau, N. J., Haur, L. ., Awang, A. Z., and Omar, W. (2016). “Repair and rehabilitation of concrete structures using confinement : A review.” Construction and Building Materials, 133, 502–515.
14. Michels, J., Widmann, R., Czaderski, C., Allahvirdizadeh, R., and Motavalli, M. (2015). “Glass transition evaluation of commercially available epoxy resins used for civil engineering applications.” Composites Part B: Engineering, 77, 484–493.

15. National Oceanic and Atmospheric Administration. (2020). "National Weather Service." <https://www.nws.noaa.gov/climate.php/xmacis.php%3Fwfo=pafc>.
16. Reay, J. T., and Pantelides, C. P. (2006). "Long-Term Durability of State Street Bridge on Interstate 80." Journal of Bridge Engineering, 11(2), 205–216.
17. Rezig, A., Nguyen, T., Martin, D., Sung, L., Gu, X., Jasmin, J., and Martin, J. W. (2006). "Relationship between chemical degradation and thickness loss of an amine-cured epoxy coating exposed to different UV environments." Journal of Coatings Technology and Research, 3(3), 173–184.
18. Sheikh, S. A., and Tam, S. (2007). "Effect of freeze-thaw climatic conditions on long-term durability of FRP strengthening systems." Ministry of Transportation of Ontario. HIIFP-037.
19. Steckel, G. L., and Hawkins, G. F. (2005). The Application of Qualification Testing, Field Testing , and Accelerated Testing for Estimating Long-Term Durability of Composite Materials for Caltrans Applications. Engineering and Technology Group The Aerospace Corporation, El Segundo,CA.
20. StEER : Structural Extreme Event Reconnaissance Network & Earthquake Engineering Research Institute (EERI) (2018). "Alaska Earthquake Preliminary Virtual Assessment Team (P-VAT) Joint Report".
21. Tatar, J., and Hamilton, H. R. (2014). "Comparison of laboratory and field environmental conditioning on FRP-concrete bond durability." Construction and Building Materials, 122, 525–536.
22. Tatar, J., Wagner, D., and Hamilton, H. R. (2016). "Structural testing and dissection of carbon fiber-reinforced polymer-repaired bridge girders taken out of service." ACI Structural Journal, 113(6).
23. Zhang, Y., Adams, R. D., and Da Silva, L. F. M. (2014). "Effects of curing cycle and thermal history on the glass transition temperature of adhesives." Journal of Adhesion, 90(4), 327–345.

The 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020)
COEX, Seoul, Korea | July 20 - 24, 2020

Paper ID ABS-0446 ISBN: …………………….

Quit Blaming ASHRAE Standard 62.1 for 1000 ppm CO_2

Andrew Persily[1,*]

[1] National Institute of Standards and Technology, Gaithersburg, USA

[*]*Corresponding email: andyp@nist.gov*

Keywords: ASHRAE, carbon dioxide; indoor air quality; metrics; standards, ventilation

1 Introduction

Indoor concentrations of carbon dioxide (CO_2) have been widely promoted as metrics of indoor air quality (IAQ) and ventilation, in many cases without a sound explanation of what they are intended to characterize or an adequate discussion of the specific application and any limitations. Many practitioners and researchers use 1800 mg/m^3 (roughly 1000 ppm$_v$) as a criteria for defining good IAQ and cite ASHRAE Standard 62.1 (ASHRAE, 2019) as the source of this value. Standard 62.1 has not contained an indoor CO_2 limit for almost 30 years, and no current ASHRAE standard contains an indoor CO_2 limit. The CO_2 limit was removed from Standard 62.1 based on the confusion that it caused and the fact that it is not a good indicator of ventilation or IAQ. Numerous papers, presentations and workshops have attempted to clarify the significance of indoor CO_2 concentrations and even advocated that they not be used as IAQ or ventilation metrics. However, these efforts have not ended the confusion, and the attribution of a 1800 mg/m^3 limit to Standard 62.1 continues. This paper describes what Standard 62.1 says about CO_2 now, what it has said in the past, explains the basis for the 1800 mg/m^3 value, and stresses that the use of a CO_2 reference value to characterize ventilation rates must consider the building type and its occupancy.

2 Historical Background

Indoor CO_2 concentrations have been discussed in the context of IAQ and ventilation for centuries. Those discussions have considered the importance of CO_2 in relation to bioeffluent perception, its application as an IAQ metric, CO_2 as a contaminant in and of itself, and its use as a tracer gas to estimate outdoor air ventilation rates. Despite many attempts to clarify the application of CO_2 to IAQ and ventilation (Persily, 1997; ASTM, 2018), much confusion has existed over the past decades and continues today. For example, there are numerous statements to the effect that a building has good IAQ because it complies with the 1000 ppm$_v$ CO_2 limit in ASHRAE Standard 62.1. This statement has multiple problems: we are not able to define good IAQ; CO_2 is not a critically important contaminant in indoor air; and, there is no 1000 ppm$_v$ limit in the standard.

ASHRAE Standard 62-1981 introduced the Indoor Air Quality Procedure, an alternative, performance-based design approach in which the ventilation system is designed to achieve target levels of indoor contaminants. This approach is in contrast to the prescriptive Ventilation Rate Procedure, in which the design must meet specific outdoor air ventilation requirements that are specific to a type of space. As part of the IAQ Procedure, the 1981 standard included a list of 20 compounds or classes of compounds with concentration limits for five of them: CO_2, chlordane, formaldehyde, ozone and radon. All of the limits were linked to a U.S. or other national government reference with the exception of CO_2. The CO_2 limit of 4500 mg/m^3 is discussed in an appendix to the 1981 standard, which noted (without reference) that 0.5 % CO_2 is a good limit based on concerns about headaches and loss of judgment. A safety factor of two is then used to account for variations in individual activity, diet and health, leading to the stated limit of 0.25 % (about 4500 mg/m^3). The 1989 standard contained concentration limits for four contaminants (CO_2, chlordane, ozone and radon) for use with the IAQ Procedure. The CO_2 limit in the 1989 standard was 1800 mg/m^3 (roughly 1000 ppm$_v$), 60 % lower than the value in the 1981 standard, but no explanation was provided for this reduction.

3 More Recent Versions of Standard 62.1

Subsequent versions of Standard 62 in 1999 and 2001 retained the contaminant limits that were in the 1989 standard, although CO_2 was removed from the table in 1999. That table was removed entirely from the 2004 version of the standard, with all discussions of contaminant limits contained in informative appendices.

The confusion regarding CO_2 in Standard 62.1 is likely associated with an informative appendix (not officially part of the standard) that was added in 1989. That appendix explained the connection between per person outdoor air ventilation rates and steady-state levels of CO_2. That discussion notes that for specified values of CO_2 generation by a person and of the outdoor CO_2 concentration, a ventilation rate of 7.5 L/s (15 cfm) per person will lead to a steady-state CO_2 concentration of 1000 ppm$_v$. That discussion was apparently interpreted by some as justifying the 1000 ppm$_v$ limit in the body of the standard under the IAQ Procedure, but that is not what the standard stated.

As noted above, the 1000 ppm$_v$ limit was removed from the standard in 1999, and the appendix was modified to better explain the connection between CO_2 concentrations and bioeffluent perception. The modified appendix noted that 7.5 L/s of outdoor air will dilute bioeffluent odors such that about 80 % of unadapted persons (visitors) are satisfied in their perception of those odors. It again noted that for assumed values of CO_2 generation, 7.5 L/s will lead to a steady-state CO_2 concentration that is 700 ppm$_v$ above outdoors. That explanation, which is not a CO_2 concentration limit, remained in the standard through 2016 and was removed from the 2019 standard.

It is important to understand that the relationship of 7.5 L/s and 1000 ppm$_v$ is only relevant to spaces for which 7.5 L/s is the outdoor air ventilation requirement. While office spaces are required to provide about 7.5 L/s per person (depending on occupant density), other spaces have ventilation requirements ranging from less than 3 L/s to 12 L/s or more. In those cases, the steady-state CO_2 concentration will be quite different from 1000 ppm$_v$, ranging from roughly 700 ppm$_v$ to 5000 ppm$_v$, again depending on the occupancy density. Therefore, identifying relevant CO_2 concentrations that correspond to ventilation rate requirements must consider the building type and its occupancy.

4 Conclusions

Despite the fact that ASHRAE Standard 62.1 has not contained an indoor CO_2 concentration limit for the past 30 years, there are many instances in which practitioners and researchers make claims that a building has good IAQ because it complies with the 1000 ppm$_v$ CO_2 limit in the standard. More recent versions of the standard do not include any statement implying that 1000 ppm$_v$ is a guideline or target value. While the direct impacts of indoor CO_2 concentrations on human health, comfort and performance are of interest, and new research is being conducted to examine those impacts, there is not yet sufficient justification to change existing ventilation standards (Fisk et al., 2019).

5 References

ASHRAE. 2019. ANSI/ASHRAE Standard 62.1-2019, Ventilation for Acceptable Indoor Air Quality. American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc., Atlanta (USA).

ASTM. 2018. ASTM 6245-2018, Standard Guide for Using Indoor Carbon Dioxide Concentrations to Evaluate Indoor Air Quality and Ventilation. ASTM International, West Conshohocken (USA).

Fisk W., Wargocki P. and Zhang X. 2019. Do Indoor CO_2 Levels Directly Affect Perceived Air Quality, Health, or Work Performance. ASHRAE Journal, 61 (9): 70-77.

Persily A. 1997. Evaluating Building IAQ and Ventilation with Indoor Carbon Dioxide ASHRAE Transactions, 103(2): 193-204.

Persily A. 2015. Challenges in developing ventilation and indoor air quality standards: The story of ASHRAE Standard 62. Building and Environment, 91: 61-69.

The 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020)
COEX, Seoul, Korea | July 20 - 24, 2020

Paper ID ABS-0430

Benchmarking Thermal Comfort Performance of Two Residential Air Distribution Systems in a Low-Load Home

Hyojin Kim[1,*], Khiem Nguyen[2], Lisa Ng[3], Brian Dougherty[3], Vance Payne[3]

[1] New Jersey Institute of Technology (NJIT), Newark, NJ, USA
[2] The Catholic University of America, Washington, DC, USA
[3] National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

[*]*Corresponding email: hyojin.kim@njit.edu*

Keywords: Small duct high velocity; Net-zero; House; Temperature; Thermal uniformity

1 Introduction

Despite broad recognition that air distribution plays an important role in thermal comfort for residential buildings, few studies have addressed the fundamental ability of residential air distribution systems to produce and deliver the selected setpoint temperature throughout a house over time. To address this deficiency, this paper compares the long-term thermal comfort performance of two air distribution systems using multiple benchmarks. The two systems, a Conventionally-Ducted Heat Pump (CDHP) and a Small Duct High Velocity (SDHV) heat pump, were used to condition the same test house, the Net-Zero Energy Residential Test Facility (NZERTF) that is located on the campus of the National Institute of Standards and Technology (NIST) in Gaithersburg, MD, USA.

2 Methods

This study analysed one-year of high resolution (i.e., 10-sec and 1-min interval) data that were collected at the NZERTF from September 2016 to August 2017. The NZERTF is a detached, single-family home that serves as a laboratory with simulated occupancy and scheduled internal loads (Pettit et al. 2015). The house provides a unique platform for comprehensive, accurate measurements to explore various designs, technologies, and control strategies to achieve net-zero energy performance. During the analysis period, the CDHP and the SDHV were operated alternately every other week in order to compare the two systems under similar weather conditions. Kim et al. (2019) provides more details on the tested systems and thermal comfort and system performance data collection.

The NZERTF results were benchmarked using metrics applied to other houses and air distribution systems, including the temperature deviation from the setpoint temperature, room-to-room temperature difference, cyclic discomfort, as well as the horizontal and vertical thermal stratification within a single room. The results were also compared against benchmarks in Air Conditioning Contractors of America (ACCA) Manual RS (1997) and ASHRAE Standard 55 (2017).

3 Results and Discussion

The results revealed the seasonal performance of the tested air distribution systems and the potential importance of data monitoring in the rooms that are not inhabited but are thermally important due to possible heat transfer from/to the primary rooms. For example, during the cooling season, the SDHV maintained better thermal uniformity than the CDHP (Figure 1). The average room-to-room temperature difference with the SDHV was 1.3 °C, which was lower than the ACCA average

Kim, Hyojin; Nguyen, Khiem; Ng, Lisa; Dougherty, Brian; Payne, Vance (Wm.). "Benchmarking Thermal Comfort Performance of Two Residential Air Distribution Systems in a Low-Load Home." Presented at the 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020), Seoul, KR. July 20, 2020 - July 24, 2020.

benchmark for cooling (1.7 °C). By comparison, the CDHP configuration resulted in an average temperature difference of 2.0 °C, and 0.8 % of the measurements exceeded the ACCA maximum benchmark for cooling (3.3 °C). When compared to 36 high-performance occupied houses in a hot and humid climate, as reported by Poerschke and Beach (2016), the SDHV also showed better thermal uniformity than that dataset based on the cumulative data above the 40th percentile in Figure 1, while the CDHP showed less uniformity.

However, for the heating season, the SDHV showed larger room-to-room temperature differences than the CDHP. The average room-to-room temperature difference of the SDHV was 1.3 °C, which exceeded the ACCA average benchmarks for heating (1.1 °C), while the CDHP had an average temperature difference of 1.1 °C. The observed seasonal differences between the two systems in whole-house thermal uniformity performance (i.e., room-to-room temperature difference) are partly attributable to the upper floor SDHV supply ductwork being entirely housed in the passively-conditioned attic, which resulted in cooler attic and second-floor temperatures during the cooling season but warmer attic and second-floor temperatures during the heating season.

The vertical and horizontal temperature stratification within a single room was analyzed using data collected from a 3 x 3 x 3 grid in a second floor bedroom. No significant amounts of horizontal and vertical stratification were observed for either system during the cooling season. However, during the heating season, the SDHV exhibited vertical stratification across the measurement grid, with the average air velocities being lower than the ACCA minimum (0.08 m/s), which indicates insufficient air circulation. However, the maximum vertical temperature difference of the SDHV (1.3 °C) was still significantly below the ASHRAE Standard 55 limit (4 °C for standing occupants).

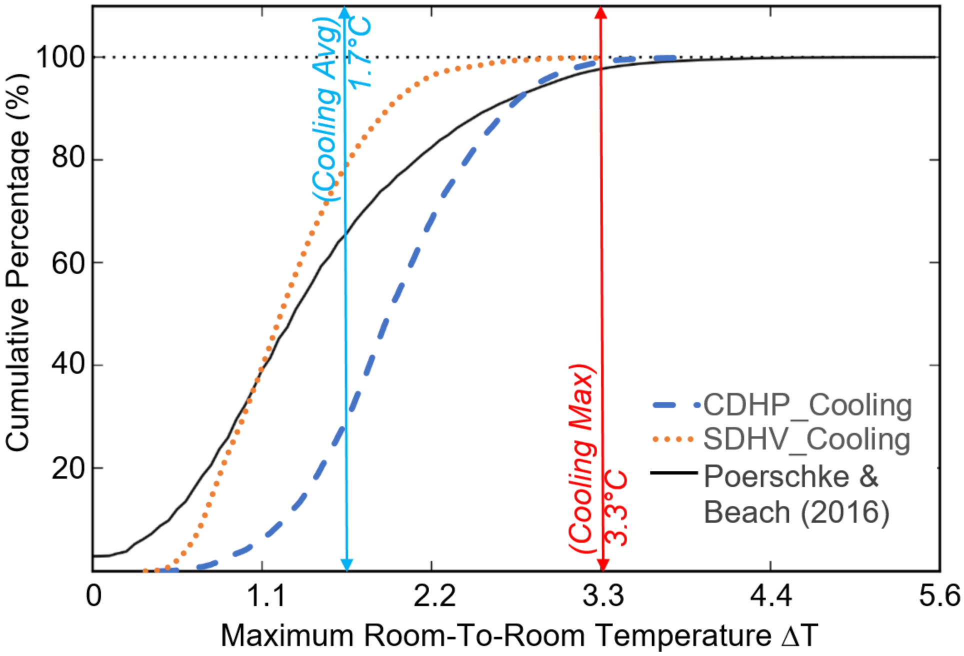


Figure 1: A Graphical Comparison of the Room-To-Room Temperature Differences.

4 Conclusions

This paper presents the results of benchmarking residential thermal comfort performance to study the HVAC system's ability to provide and maintain uniform space temperatures throughout the house. The ACCA Manual RS's primary focus on air temperature provides reasonable benchmarks for this study, since most residential HVAC systems are single-zone systems that are configured to control the thermal conditions of the house solely based on air temperature at the thermostat. There would be value in re-examining the ACCA Manual RS benchmarks, which were developed in 1997 based on HVAC systems and homes of 20 years ago, using more recent house and systems and more representative datasets for benchmarking.

5 Acknowledgement

NJIT and Catholic University participation in this study was funded through the NIST Measurement Science and Engineering (MSE) Research Grant Program.

6 References

ACCA. 1997. ACCA Manual RS - Comfort, Air Quality, and Efficiency by Design. Arlington, VA: Air Conditioning Contractors of America, Inc.

ASHRAE. 2017. ANSI/ASHRAE Standard 55-2017, Thermal Environmental Conditions for Human Occupancy. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

Kim, H., K. Nguyen, A. McGuinness, and T.V. Dai. 2019. Characterization of Residential Air Distribution System Performance for Thermal Comfort. NIST GCR 19-021.

Pettit, B., C. Cates, A. Fanney, and W. Healy. 2015. Design Challenges of the NIST Net Zero Energy Residential Test Facility. Gaithersburg, MD: NIST Technical Note 1847.

Poerschke, A. and R. Beach. 2016. Comfort in High-Performance Homes in a Hot-Humid Climate. U.S. DOE Building America Subcontract Report DOE/GO-102016-4762. Pittsburgh, PA: Integrated Building and Construction Solutions (IBACOS).

The 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020)
COEX, Seoul, Korea | July 20 - 24, 2020

Paper ID ABS-0433

Characterization of Long-Term Sub-Hourly Thermal Comfort Performance Data

Hyojin Kim[1,*], Lisa Ng[2], Brian Dougherty[2], Vance Payne[2]

[1] New Jersey Institute of Technology (NJIT), Newark, NJ, USA
[2] National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

[*]*Corresponding email: hyojin.kim@njit.edu*

Keywords: Architectural plan trend animations; Data decomposition; Weather-normalized thermal comfort

1 Introduction

Current thermal comfort standards do not provide guidelines on how to graphically characterize and benchmark long-term, sub-hourly, thermal comfort. In an effort to address this need, statistical characterizations and visualization methods are demonstrated for quantifying and understanding the long-term thermal comfort in single-family homes. One year of sub-hourly field data were used for this study. The data corresponds to times when two air-source heat pump (HP) systems were alternately used to condition the same well-insulated, well-sealed residence, the Net-Zero Energy Residential Test Facility (NZERTF) on the campus of the National Institute of Standards and Technology (NIST) in Gaithersburg, MD, USA. Kim et al. (2019) provides more details on the house, HP systems, and thermal comfort and system performance data collection.

2 Methods

To extract meaningful information from large datasets, the proposed statistical characterization methods decomposed the field data by the type of HP system (i.e., a Conventionally-Ducted Heat Pump (CDHP) and a Small Duct High Velocity (SDHV) heat pump), ON versus OFF cycles, and season. A modified box and whisker plot was proposed and used to characterize extreme variations, while focusing on the measurements falling in the extreme percentiles: ±1.5 % (i.e., 0 % to 1.5 % and 98.5 % to 100 %), ±2.5 %, ±5 %, and ±10 %. These four percentiles are consistent with the recommended criteria for acceptable deviations (i.e., 3% and 5%) in Annex G of the European standard, DIN EN 15251 (2007). To better understand the temporal variations of the long-term thermal comfort data, this study also explored advanced characterizations, including animation techniques, which accounted for the effects of outdoor climate and the time of the day.

3 Results and Discussion

Partitioning data between HP system ON versus OFF provided useful insights on how the two systems responded to different thermal needs. Similarly, identifying and separately grouping days when the HP systems provided little or no space conditioning helped the interpretation process. These separately-grouped days, loosely defined as being part of the transitional seasons, occurred because the thermostat was set to either COOL with a 23.9 °C setpoint or HEAT with a 21.1 °C setpoint, thus creating an unconditioned temperature or "float" zone between these setpoints. Also, because the auto changeover feature of the thermostat was disabled, cooling (or heating) may be prevented on a day where the weather changed appreciably but the thermostat was still set to HEAT (or COOL).

Kim, Hyojin; Ng, Lisa; Dougherty, Brian; Payne, Vance (Wm.). "Characterization of Long-Term Sub-Hourly Thermal Comfort Performance Data." Presented at the 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020), Seoul, KR. July 20, 2020 - July 24, 2020.

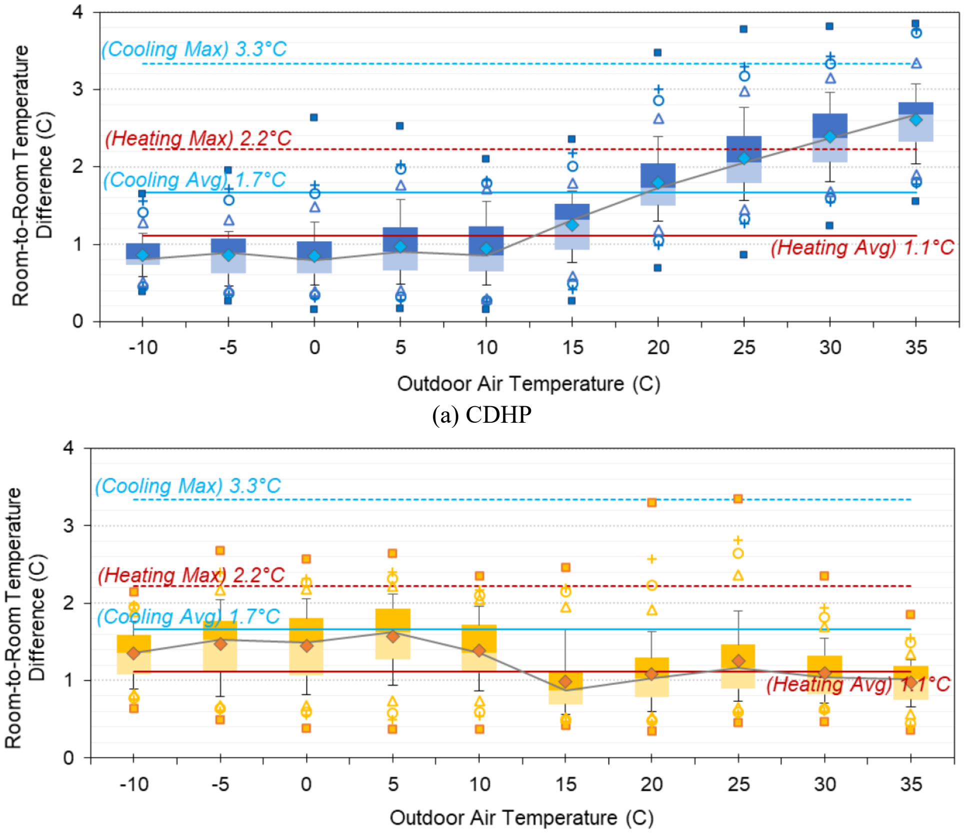
Characterization of temporal variations in room temperatures relative to coincident outdoor ambient temperature measurements allowed a weather-normalized comparison of the thermal comfort delivered by the two HP systems. When the CDHP was in operation, the room-to-room temperature difference had a strong association with the outdoor temperature during the cooling season. By comparison, no comparable relationship was observed with the SDHV (Figure 1). This behavior was affected by the different locations of the supply ductwork for the two HP systems. The SDHV supply ductwork is located within the passively-conditioned attic, with ceiling supply diffusers in the upstairs rooms. These upstairs rooms were conditioned using the high wall and floor supply registers of the CDHP system.

Under the same outdoor dew point temperatures, the SDHV maintained the rooms drier when the systems cycled ON, which indicates better dehumidification performance than the CDHP. However, when the systems cycled OFF, the SDHV had higher humidity ratios at the higher outdoor dew point temperatures, with a number of incidences exceeding the upper recommended humidity ratio limit of the ASHRAE Standard 55 (2017), 0.012 kg H_2O/kg dry air. This occurred because the SDHV turned off every day at midnight for an extended period of time, causing humidity accumulation.

Time-of-day characterization was useful to understand the dynamic interactions between thermal conditions and changes in internal heat gains from occupants, lighting, and plug loads For example, the living room temperature was always lower at night because neither occupants nor plug loads were simulated to exist at that time. However, the opposite pattern (i.e., lower temperatures during the daytime due to no internal loads) was observed in the master bedroom on the second floor.

Lastly, the animated analysis obtained by mapping temperatures on the 3-D floor plans was helpful to quickly relate the measurements with the geometry, orientation, or space function over the course of a day or month. They were also useful to identify any anomalies in the data at a single glance. The proposed visualizations were one of the ways for efficiently evaluating the voluminous dataset.

Kim, Hyojin; Ng, Lisa; Dougherty, Brian; Payne, Vance (Wm.). "Characterization of Long-Term Sub-Hourly Thermal Comfort Performance Data." Presented at the 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020), Seoul, KR. July 20, 2020 - July 24, 2020.



(a) CDHP

(b) SDHV

Figure 1: Binned Room-To-Room Temperature Differences Against Outdoor Temperatures.

4 Conclusions

A long-term measurement of high-frequency thermal comfort is difficult due to challenges in data collection, processing, and inspection. There is a need for effective methods to translate the large volume of data into useful information, which can also be useful for benchmarking. The proposed, modified box and whisker plot was informative when statistically characterizing one-year granular thermal comfort data related to outdoor climate and the time of the day. This characterization should be performed not only for the primary rooms but also for other rooms that are thermally important, such as the attic and basement, due to possible heat transfer from/to the primary rooms.

5 Acknowledgement

The involvement of NJIT in this study was funded through the NIST Measurement Science and Engineering (MSE) Research Grant Program.

6 References

ASHRAE. 2017. ANSI/ASHRAE Standard 55-2017, Thermal Environmental Conditions for Human Occupancy. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

CEN. 2007. EN 15251:2007, Indoor Environmental Input Parameters for Design and Assessment of Energy Performance of Buildings Addressing Indoor Air Quality, Thermal Environment, Lighting and Acoustics. Brussels, Belgium: European Committee for Standardization.

Kim, H., K. Nguyen, A. McGuinness, and T.V. Dai. 2019. Characterization of Residential Air Distribution System Performance for Thermal Comfort. NIST GCR 19-021. 287 pages (December).

Kim, Hyojin; Ng, Lisa; Dougherty, Brian; Payne, Vance (Wm.). "Characterization of Long-Term Sub-Hourly Thermal Comfort Performance Data." Presented at the 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020), Seoul, KR. July 20, 2020 - July 24, 2020.

Classifying Memory Bugs Using Bugs Framework Approach

Irena Bojanova
SSD, ITL
NIST
Gaithersburg, MD, USA
irena.bojanova@nist.gov

Carlos Eduardo Galhardo
Dimel, Sinst
INMETRO
Duque de Caxias, RJ, Brazil
cegalhardo@inmetro.gov.br

***Abstract*—In this work, we present an orthogonal classification of memory corruption bugs, allowing precise structured descriptions of related software vulnerabilities. The Common Weakness Enumeration (CWE) is a well-known and used list of software weaknesses. However, it's exhaustive list approach is prone to gaps and overlaps in coverage. Instead, we utilize the Bugs Framework (BF) approach to define language-independent classes that cover all possible kinds of memory corruption bugs. Each class is a taxonomic category of a weakness type, defined by sets of operations, cause→consequence relations, and attributes. A BF description of a bug or a weakness is an instance of a taxonomic BF class, with one operation, one cause, one consequence, and their attributes. Any memory vulnerability then can be described as a chain of such instances and their consequence–cause transitions. We showcase that BF is a classification system that extends the CWE, providing a structured way to precisely describe real world vulnerabilities. It allows clear communication about software bugs and weaknesses and can help identify exploit mitigation techniques.**

***Keywords*—Bug classification, bug taxonomy, software vulnerability, software weakness, memory corruption.**

I. Introduction

Software bugs in memory allocation, use, and deallocation may lead to memory corruption and memory disclosure, opening doors for cyberattacks. Classifying them would allow precise communication and help us teach about them, understand and identify them, and avoid security failures. For that, we utilize the Bug Framework (BF) approach [1].

The Common Weakness Enumeration (CWE) [2] and the Common Vulnerabilities and Exposures (CVE) [3] are well-known and used lists of software security weaknesses and vulnerabilities. However, the CWE's exhaustive list approach is prone to having gaps and overlaps in coverage, as demonstrated by the National Vulnerability Database (NVD) effort to link CVEs to appropriate CWEs [4]. Instead, we utilize the BF approach to define four language-independent, orthogonal classes that cover all possible kinds of memory related software bugs and weaknesses: Memory Allocation Bugs (MAL), Memory Use Bugs (MUS), Memory Deallocation Bugs (MDL), and Memory Addressing Bugs (MAD). This BF Memory Bugs taxonomy can be viewed as a structured extension to the memory-related CWEs, allowing bug reporting tools to produce more detailed, precise, and unambiguous descriptions of identified memory bugs.

In this paper, we first summarize the latest BF approach and methodology. Next, we analyze the types of memory corruption bugs and define the BF Memory Bugs Model. Then, we present our BF memory bugs classes and showcase they provide a better, structured way to describe CVE entries [3]. We identify the corresponding clusters of memory corruption CWEs and their relations to the BF classes. Finally, we discuss the use of these new BF classes for identifying exploit mitigation techniques.

II. BF Approach and Methodology

BF's approach is different from CWE's exhaustive list approach. BF is a classification. Each BF class is a taxonomic category of a weakness type. It relates to a distinct phase of software execution, the operations specific for that phase and the operands required as input to those operations.

We define a software bug as a coding error that needs to be fixed. A weakness is caused by a bug or ill-formed data. A weakness type is also a meaningful notion, as different vulnerabilities may have the same type of underlying weaknesses. We define a vulnerability as an instance of a weakness type that leads to a security failure. It may have more than one underlying weaknesses linked by causality.

BF describes a bug or a weakness as an improper state and its transition. The transition is to another weakness or to a failure. An improper state is defined by the tuple ($\texttt{operation}$, $\texttt{operand}_1$, $\cdots$, $\texttt{operand}_n$), where at least one element is improper. The initial state is always caused by a bug; a coding error within the operation, which if fixed will resolve the vulnerability. An intermediate state is caused by ill-formed data; it has at least one improper operand. Rarely an intermediate state may also have a bug, which if fixed will also resolve the vulnerability. The final state, the failure, is caused by a final error (undefined or exploitable system behavior), which usually directly relates to a CWE [2]. A transition is the result of the operation over the operands.

BF describes a vulnerability as a chain of improper states and their transitions. Each improper state is an instance of a BF class. The transition from the initial state is by improper operation over proper operands. The transitions from intermediate states are by proper operations with at least one improper operand.

In some cases, several vulnerabilities have to be present for an exploit to be harmful. The final errors resulting from different chains converge to cause a failure. The bug in at least one of the chains must be fixed to avoid that failure.

We call a BF class the set of operations, the valid cause→consequence relations for these operations, their at-

Disclaimer: Certain trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor that they are necessarily the best available for the purpose.

Bojanova, Irena. "Classifying Memory Bugs Using Bugs Framework Approach." Presented at 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC 2021). July 12, 2021 - July 16, 2021.

tributes, and sites. The attributes are qualifiers for the operations and the operands that help understand how severe a bug is. The sites show where in code a bug might occur. The BF classes are orthogonal by design; their sets of operations do not overlap.

The taxonomy of a particular bug or weakness is based on one BF class. Its description is an instance of a taxonomic BF class with one cause, one operation, one consequence, and their attributes. The operation binds the cause→consequence relation – e.g., deallocation via a dangling pointer leads to a final error known as double free [5].

The methodology for developing a BF class is as follows: (1) Identify the phase specific for a kind of bugs. (2) Identify the operations for that phase. (3) Define a BF bugs model showing operations flow. (4) Identify all causes. (5) Identify all consequences that propagate as a cause to a next weakness. (6) Identify all consequences that are final errors. (7) Identify attributes useful to describe such a bug/weakness. (8) Identify possible sites in code.

III. Memory Bugs Model

Each memory related bug or weakness involves one memory operation. Each *operation* is over a region of memory or over the address needed to reach it. That memory is used for storing data and has an important property: it is finite. It has *boundaries* and it has *size*. We call this piece of memory, with a well-defined size, an *object*. It is used to store a primitive data or a data structure. The memory address should be held by at least one *pointer* or determined as an offset on the stack, otherwise the object will be unreachable. The object and the pointer are the operands of that memory operation (see definitions in Table III).

Memory bugs could be introduced at any of the phases of an object's lifecycle: *address formation*, *allocation*, *use*, and *deallocation*. The BF Memory Bugs Model helps identify where in these phases bugs could occur (Fig. 1). The phases correspond to the BF memory bugs classes: Memory Addressing Bugs (MAD), Memory Allocation Bugs (MAL), Memory Use Bugs (MUS), and Memory Deallocation Bugs (MDL). All possible memory operations are grouped by phase. The presented operations flow helps in identifying possible chains of bugs/weaknesses.

The operations under MAD (Fig. 1) are on forming or modifying a pointer: *Initialize*, *Reposition*, and *Reassign*. Bugs in pointer initialization could result in pointers to meaningless objects. Moving a pointer via a bugged Reposition could get it pointing outside the object bounds. Bugs in Reassign could connect a pointer to a wrong object. See definitions of MAD operations in Table Ia.

The operations under MUS are on reading or writing the content of an object through one of its pointers: *Initialize*, *Read*, *Write*, *Clear*, and *Dereference*. Bugs in object initialization could lead to use of random or malicious data. Bugs in write could alter data wrongly. Bugs in Clear could leak confidential information such as passwords and cryptographic private keys. Bugs in Dereference are practically unsuccessful reading or unsuccessful writing. See definitions of MUS operations in Table Ic.

The operations under MAL are on creating an object or extending it through one of its pointers: *Allocate*, *Extend*, and *Reallocate–Extend* (see definitions in Table Ib). The operations under MDL are on destroying or reducing an object through one of its pointer: *Deallocate*, *Reduce*, and *Reallocate–Reduce* (see definitions in Table Id). Both MAL and MDL operations affect the boundaries and the size of the object. Bugs in Reallocate may concern multiple pointers to the same object. Allocation in excess or failure to deallocate unused objects could exhaust memory. Excessive reduction of allocated memory could lead to an object that is too little for the data it needs to store.

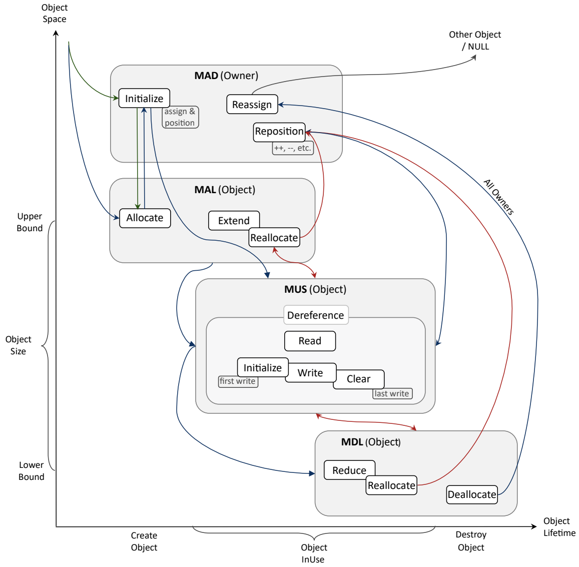


Fig. 1: The BF Memory Bugs Model. Comprises four phases, corresponding to the BF classes MAD, MAL, MUS, and MDL. Shows the memory operations flow: blue arrows – the main flow; green arrows – flow for allocation at a specific address; red – extra flow in case of reallocation.

The possible flow between operations from different phases is depicted on Fig. 1 with colored arrows: blue is for the main flow; green is for allocation requested at a specific address; red is for the extra flow in case of reallocation.

Following the blue arrows, the very first operation is MAL Allocate an object. Following the green arrows, the first operation is MAD Initialize a pointer. Next operation, following the blue arrows, should be MAD Initialize the pointer to the address returned by Allocate. While, following the green arrows, next operation should be MAL Allocate an object at the address the pointer holds.

After an object is allocated and its pointer is initialized, it can be used via MUS Read or MUS Write. The boundaries and the size of an object are set at allocation, then they can be changed by any MAL or MDL operation.

If an object is owned by more than one pointer, Reallocate (in MAL or MDL) should be followed by Reposition over all these owners. A Deallocate an object operation should properly be followed by Reassign of all its pointers to either `NULL` or another object.

IV. BF Memory Bugs Classes

We define the BF Memory Bugs classes as follows:

TABLE I: Operations

(a) MAD (Memory Addressing)

Operation Value	Definition
Initialize (pointer)	The first assign of an object address to a pointer; positions the pointer at the start of the object.
Reposition	Changes the pointer to another position inside its object.
Reassign	Changes the pointer to a different object.

(b) MAL (Memory Allocation)

Operation Value	Definition
Allocate	Reserves space in memory for an object; defines its initial boundaries and size.
Extend	Allocates additional memory for an object in the same space; redefines its boundaries and size.
Reallocate–Extend	Allocates a new larger piece of memory for an object at a new address, copies the object content there, reassigns its pointer, and deallocates the previous piece of memory.

(c) MUS (Memory Use)

Operation Value	Definition
Initialize (object)	The first write into an object, after it is allocated.
Read	Gets content from an object.
Write	Puts content into an object.
Clear	The very last write into an object, before it is deallocated.
Dereference	Overreaches Initialize, Read, Write, and Clear, focus is on object access, no matter if it's for reading or for writing.

(d) MDL (Memory Deallocation)

Operation Value	Definition
Deallocate	Releases the allocated memory of an object.
Reduce	Deallocates part of the object memory; redefines its boundaries and size.
Reallocate–Reduce	Allocates a new smaller space in memory for an object at a new address, copies part of the object content there, reassigns the pointer, and deallocates the previous piece of memory.

Memory Addressing Bugs (MAD) – *The pointer to an object is initialized, repositioned, or reassigned to an improper memory address.*

Memory Allocation Bugs (MAL) – *An object is allocated, extended, or reallocated (while extending) improperly.*

Memory Use Bugs (MUS) – *An object is initialized, read, written, or cleared improperly.*

Memory Deallocation Bugs (MDL) – *An object is deallocated, reduced, or reallocated (while reducing) improperly.*

Each of these classes represents a phase, aligned with the Memory Bugs Model, and is comprised of sets of operations, cause→consequence relations, and attributes. Fig. 2, Fig. 3, Fig. 4, and Fig. 5 show the specific sets for memory addressing, allocation, use, and deallocation bugs, respectively. Only the values listed on the corresponding figure should be used to describe that kind of bugs or weaknesses.

A. Operations

All BF classes are being designed to be orthogonal; their sets of operations should not overlap. The operations in which memory bugs could happen are defined in Table I.

The MAD operations are: Initialize (Pointer), Reassign, Reposition. They reflect improper formation of an address.

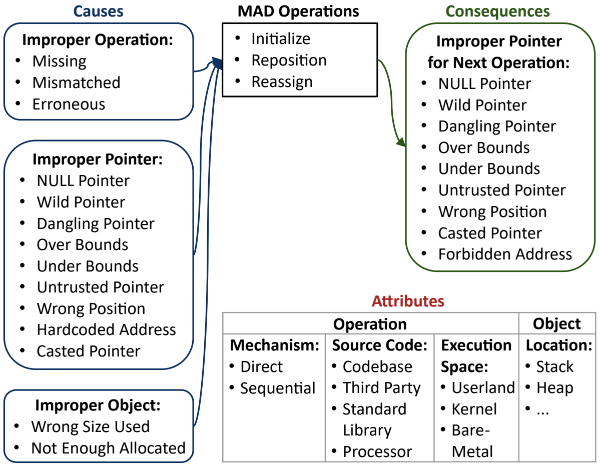


Fig. 2: The Memory Addressing Bugs (MAD) class.

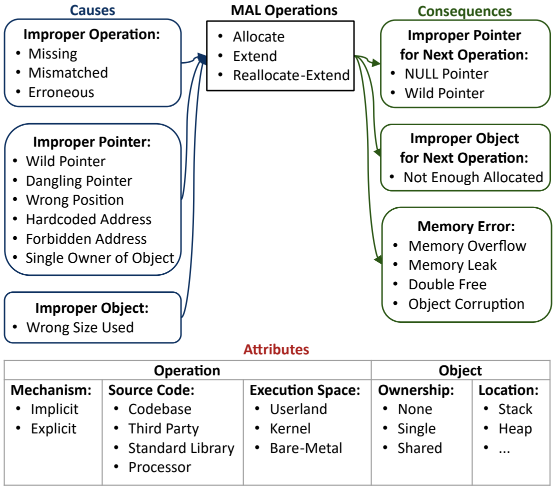


Fig. 3: The Memory Allocation Bugs (MAL) class.

The MAL operations are: Allocate, Extend, and Reallocate–Extend. They reflect improper formation of an object. The MUS operations are: Initialize (Object), Dereference, Read, Write, Clear. They reflect improper use of an object. The MDL operations are: Deallocate, Reduce, Reallocate–Reduce. They reflect improper release of an object. MAD Initialize and MUS Initialize are not overlapping, as the former is about the address, the latter is about the object.

B. Causes

A cause is either an improper operation or an improper operand. The values for improper memory operations are: *Missing*, *Mismatched*, and *Erroneous*. See definitions in Table II. The operands of a memory operation are pointer and object. See definitions in Table III. All values for improper operands of a memory operation are defined in Table IV.

An improper pointer could be a reference. Comments could be used to provide details, such as the pointer or reference identifier. An improper object could be a primitive

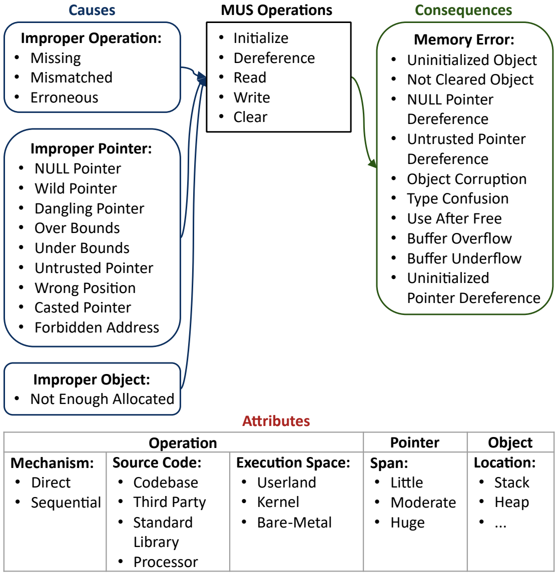


Fig. 4: The Memory Use Bugs (MUS) class.

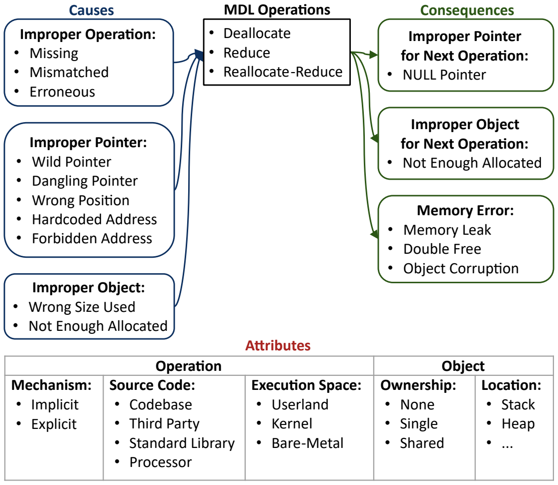


Fig. 5: The Memory Deallocation Bugs (MDL) class.

TABLE II: Improper Operations

Value	Definition	Example
Missing	The operation is absent.	Missing object initialization.
Mismatched	The deallocation function does not match the allocation function used for the same object.	Use of `free()` on an object allocated with `new`.
Erroneous	There is a bug is in the implementation of the operation.	Allocation with `malloc()` returns a non existing address.

data type or a data structure. Comments could be used to provide details, such as the object data type and identifier.

All possible causes for memory bugs are defined in Table II and Table IV. However, refer Fig. 2, Fig. 3, Fig. 4, and Fig. 5 for causes applicable to each class.

When describing a chain of bugs/weaknesses, the pointer and the object should be analyzed carefully, as they may be different for each improper state. The description should reflect the changes and provide details in the comments.

C. Consequences

A consequence is either an improper operand or a final error. As a consequence, an improper pointer or an improper object would become a cause for a next weakness. These consequence–cause transitions explain why these two kinds of consequences have the same possible values as the corresponding kinds of causes (see Table II and Table IV).

All possible memory errors are defined in Table V.

The only kind of MAD consequences is Improper Pointer, which means a MAD bug or weakness is always followed by another memory weakness, such as of MAL, MUS, or MDL. The only kind of MUS consequences is Memory Error, which means MUS always ends in a failure.

All possible consequences for memory bugs are defined in Table IV and Table V. However, refer Fig. 2, Fig. 3, Fig. 4, and Fig. 5 for consequences applicable to each class.

D. Attributes

An *attribute* provides additional useful information about the operation or its operands. All possible attributes for memory bugs are defined in Table VI.

All Memory Bugs classes have the following attributes: *Source Code*, *Execution Space*, and *Location*. They explain

TABLE III: Operands

Concept	Definition
Object	A memory region used to store data.
Pointer	A holder of the memory address of an object.

TABLE IV: Improper Operands

(a) Improper Pointer

Value	Definition
NULL Pointer	Points to the zero address, a specific invalid address.
Wild Pointer	Points to an arbitrary address, because it has not been initialized or an erroneous allocation routine is used.
Dangling Pointer	Still points to the address of its successfully deallocated object.
Over Bounds	Points over the bounds of its object.
Under Bounds	Points under the bounds of its object.
Untrusted Pointer	The pointer is modified to an improperly checked address.
Wrong Position	Points to a miscalculated position inside object bounds.
Hardcoded Address	The pointer points a wrong specific address.
Casted Pointer	The pointer does not match the type of the object, due to wrong type casting.
Forbidden Address	The pointer points to an OS protected or non-existing address.
Single Owner of Object	The only pointer of an already allocated object is used to allocate a new object.

(b) Improper Object

Value	Definition
Not Enough Allocated	The allocated memory is too little for the data it should store.
Wrong Size Used	The value used as size does not match the real size of the object.

Bojanova, Irena. "Classifying Memory Bugs Using Bugs Framework Approach." Presented at 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC 2021). July 12, 2021 - July 16, 2021.

TABLE V: Memory Errors

Value	Definition	Risk
Memory Overflow	More memory requested than available.	Stack/heap exhaustion.
Memory Leak	An object has no pointer pointing to it.	Resource exhaustion. Application crash. DoS.
Double Free	Attempt to deallocate a deallocated object or via an uninitialized pointer.	Arbitrary code execution.
Object Corruption	Object data is unintentionally altered.	Wrong/unexpected results.
Uninitialized Object	Object data is not filled in before use.	Controlled or left over data.
Not Cleared Object	Object data not overwritten before deallocation.	Information exposure (e.g. private keys).
NULL Pointer Dereference	Attempt to access an object for read or write through a NULL pointer.	Program crash. Arbitrary code execution (in some OSs).
Untrusted Pointer Dereference	Attempt to access an object via an altered pointer (not legitimate dereference of tainted pointers).	DoS. Arbitrary code execution.
Type Confusion	Pointer and object have different types.	Vtable corruption. Hijack.
Use After Free	Attempt to use a deallocated object.	Arbitrary code execution.
Buffer Overflow	Read or write above the object upper bound.	Arbitrary code execution. Information exposure.
Buffer Underflow	Read or write below the object lower bounds.	Arbitrary code execution. Information exposure.
Unitialized Pointer Derefereance	An attempt to access an object for read or write via an uninitialized pointer.	Control flow hijack.

TABLE VI: Attributes

(a) MAD, MAL, MUS, MDL Attributes

Name	Value	Definition
Source Code	Codebase	The operation is in programmer's code – in the application itself.
	Third Party	The operation is in a third party library.
	Standard Library	The operation is in the standard library for a particular programming language.
	Language Processor	The operation is in the tool that allows execution or creates executable (compiler, assembler, interpreter).
Execution Space	Userland	The bugged code runs in an environment with privilege levels, but in unprivileged mode (e.g., ring 3 in x86 architecture).
	Kernel	The bugged code runs in an environment with privilege levels with access privileged instructions (e.g., ring 0 in x86 architecture).
	Bare-Metal	The bugged code runs in an environment without privilege control. Usually, the program is the only software running and has total access to the hardware.
Location [1]	Stack	The object is a non-static local variable (defined in a function, a passed parameters, or a function return address).
	Heap	The object is a dynamically allocated data structure (e.g., via `malloc()` and `new`).

(b) MAD and MUS Attributes

Name	Value	Definition
Mechanism	Direct	The operation is performed over a particular object element.
	Sequential	The operation is performed after iterating over the object elements.

(c) MAL and MDL Attributes

Name	Value	Definition
Mechanism	Implicit	The operation is performed without a function call.
	Explicit	The operation is performed by a function/ method) call.
Ownership	None	The object has no owner.
	Single	The object has one owner.
	Shared	The object has more than one owner.

(d) MUS Attributes

Name	Value	Definition
Span	Little	A few bytes of memory are accessed.
	Moderate	Several bytes of memory are accessed, but less than 1 KB.
	Huge	More than 1 KB of memory is accessed.

where a bug is in three dimensions: where is the operation in the program, where its code is running, and where the object is stored in memory. See definitions of values in Table VIa.

All Memory Bugs classes have also the operation attribute *Mechanism*, but with different possible values.

For MAD and MUS *Mechanism* qualifies an operation as *Direct* or *Sequential*, depending on if an object element is accessed directly or after going through previous elements. See definitions of values in Table VIb.

For MAL and MDL *Mechanism* qualifies an operation as *Implicit* or *Explicit*. For MAL, *Implicit* means automatic compile-time allocation. Improper results from implicit allocation are not enough memory allocated or too much memory requested, overflowing the stack (e.g., via a recursion). For MDL, *Implicit* means automatic deallocation at the end of scope. Bugs in automatic memory allocation or deallocation are rare (e.g., the gcc compiler bug [6]). For MAL, *Explicit* means dynamic run time allocation (e.g. using `malloc()` or `new`). For MDL, *Explicit* means dynamic run time deallocation (e.g. using `free()` or `del`).

MAL and MDL have also the pointer attribute *Ownership*. It shows how many pointers point to an object: *None*, *Single*, and *Shared*. See definitions of values in Table VIc. For MAL, it shows how many pointers hold the allocated object. For MDL, if an object has no pointer pointing to it, it will be unreachable for deallocation in an environment without a garbage collector. Multiple pointers to the same object could lead to race conditions and dangling pointers.

MUS has also the pointer attribute *Span*. It shows how many bytes are being used: *Little*, *Moderate*, *Huge*, depending on if those are a few, more than a few and less than one KB, or more than one KB. See definitions of values in Table VId.

E. Sites

MAD sites are any changes to a pointer via assignment (=) or repositioning via an index (`[]`) or pointer arithmetics (e.g., `p++` and `p--`).

[1] Other proper values should be used for different kinds of memory layout. For example, Uninitialized Data Segment, Data Segment, and Code Segment layout should be added for C language layout [7].

MAL sites are any allocation routine (e.g., `malloc()`) or operator (e.g. `new`), declaration of a variable with implicit allocation, OOP constructor, or extension routine (e.g., `realoc()`) or adding elements to a container object.

MUS sites are any dereference operators in the source code (`*`, `[]`, `->`, `.`).

MDL sites are any deallocation routine (e.g., `free()`) or operator (e.g. `del`), end of scope for implicit allocated variables, OOP destructor, or reduction routine (e.g., `realoc()`) or removing elements from a container object.

V. The BF Memory Classes as CWE Extension

BF Memory Bugs taxonomy can be used by bug reporting tools, as it is a structured extension over memory-related CWEs [2]. All Memory Error consequences from the BF classes (Table V) relate to one or more CWEs.

We have generated a digraph (Fig. 6) of all memory-related CWEs to show how they correspond to the possible BF Memory Error consequences (Table V). An edge starts at a parent CWE and ends at a child CWE. The outline style of a CWE node indicates the CWE level of abstraction: pillar, class, base, or variant. Bug reporting tools would use base or variant CWEs, but they may also use higher abstraction level CWEs if there is not enough specific information about the bug or if there is no related base CWE.

The digraph helped us identify clusters of memory-related CWEs. All these CWEs can be tracked as children of the pillar CWE-664, with the only exception of CWE-476 (NULL Pointer Dereference). The largest cluster comprises CWE-118 and the children of CWE-119, which are weaknesses associated with reading and writing outside the boundaries of an object. The second cluster comprises the children of CWE-400 and CWE-665, which are mainly weaknesses related to memory allocation and object initialization. The children of CWE-404, which are weaknesses associated with improper memory cleanup and release, form the third cluster. The smallest cluster comprises CWE-704, CWE-588 and CWE-843, which are memory use or deallocation weaknesses due to the mismatch between pointer and object types.

The color of a CWE node (Fig. 6) indicates the BF memory class associated with that CWE. A BF class is associated with a CWE if the BF class has a Memory Error consequence covered by the CWE description. CWEs related to the BF MUS memory errors are presented in blue, CWEs related only to MAL are presented in pink, and CWEs related to both MAL and MDL are presented in green.

Most of the BF MUS Memory Error consequences (Fig. 4) relate to CWEs from the CWE-118 cluster. The Memory Error consequences from BF MAL and BF MDL (Fig. 3 and Fig. 5) relate to CWEs across clusters. Note that the BF MAD class (Fig. 2) has no Memory Error consequences, so it does not directly relate to any CWE.

The BF Memory Bugs model (Fig. 1) reflects the lifecycle of an object. While the pillar CWE-664 reflects the "lifetime of creation, use, and release" of a resource, it is quite broad. It is the parent of many CWEs that are not strictly memory-related. We use asterisks (*) to denote CWEs that are about any resource. CWE-704 is not a memory-related CWE, but is visualized on the digraph to show all the parent-child relationships.

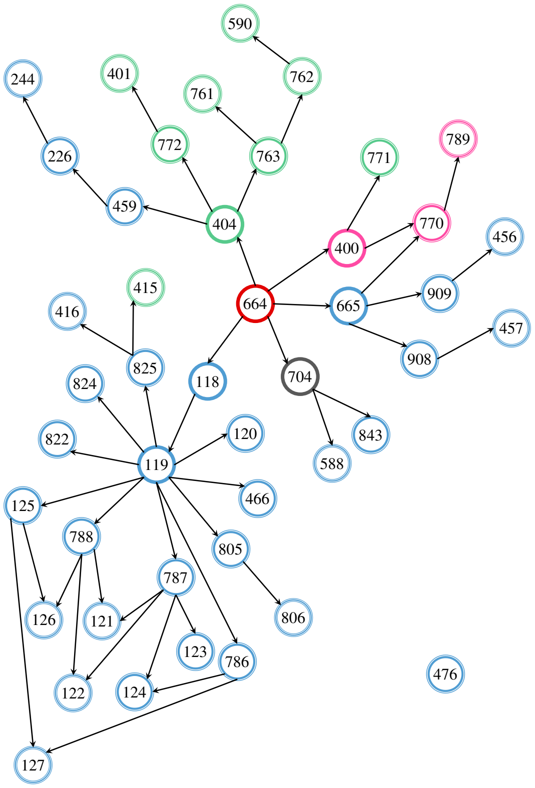


Fig. 6: A digraph of all memory related CWEs. Triple line – variant, double line – base, single line – class, and thick red line – pillar. BF class correspondence: pink is only for MAL, green is for both MDL and MAL, and blue is for MUS.
→ Click on an ID to open the CWE entry.

The identified clusters of memory CWEs do not strictly correspond to the phases of address formation, allocation, use, and deallocation. CWEs related to a phase appear in more than one cluster. In addition, CWE-118 and CWE-119 are strictly about memory but cover more than one phase.

Viewed as a structured extension, the BF Memory Bugs classes relate to CWEs through particular Memory Error consequences. For BF MAL: Memory Overflow – relates to CWEs: 400*, 770*, and 789; Memory Leak – to CWEs: 401, 404*, 771*, and 772*; Double Free – to CWE-415; Object Corruption – to CWEs: 404*, 590, 761, 762, and 763.

For BF MUS: Uninitialized Object – relates to CWEs: 456, 457, 665*, 908*, and 909*; Not Cleared Object – to CWEs: 226*, 244, and 459*; NULL Pointer Dereference – to CWE-476; Untrusted Pointer Dereference – to CWEs: 119 and 822; Type Confusion – to CWEs: 588 and 843*; Use After Free – to CWEs: 119, 416, and 825; Buffer Overflow – to CWEs: 118, 119, 120, 121, 122, 123, 125, 126, 466, 805, 806, 787, and 788; Buffer Underflow – to CWEs: 118, 119, 122, 123, 124, 125, 127, 466, 786, 787, 805, and 806; Unitialized Pointer Dereference – to CWEs: 119 and 824. There are no related CWEs to BF MUS Object Corruption.

For BF MDL: Memory Leak – relates to CWEs: 401, 404*, and 771*; Double free – to CWE-415; Object Corruption – to CWEs: 404*, 761, 762, and 763.

VI. Showcases and Discussion

In this section, we use the new BF Memory Bugs classes for precise descriptions of real world software vulnerabilities. We also provide the real world fixes of each bug.

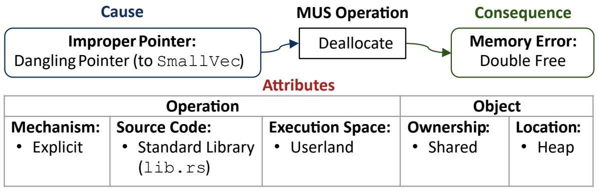
A. CVE-2018-20991 – Rust SmallVec Iterator Panic

This vulnerability is listed in CVE-2018-20991 and discussed in [8]. The source code could be found at [9]. In Rust, a panic is an unrecoverable error that terminates the thread, possibly unwinding its stack (calling destructors as if every function instantly returned) [10].

a) Brief Description: Rust is a multi-paradigm programming language focused on safe concurrency. It has a similar syntax to C++ and offers features to deal with dynamic memory allocation, such as smart pointers [11]. In general, a Rust programmer does not need to keep track of memory allocation and deallocation, as the language is designed to be memory safe this way.

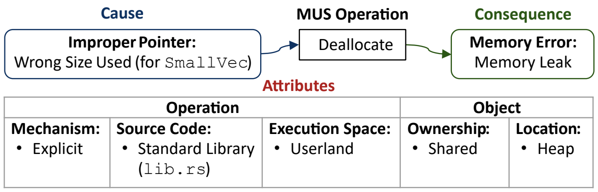
b) Analysis: The versions before Rust 0.6.3 have a bug in the `lib.rs` file. The `insert_many()` method in the `SmallVec` class has two parameters: an iterable `I` and an `index`. The method inserts all elements in the iterable `I` at position `index`, shifting all the following elements backwards. In the `SmallVec` class, if an iterator passed to `SmallVec::insert_many()` panics in `Iterator::next`, the destructor is called while the vector is in an inconsistent state, possibly causing double free (deallocation via references to same object). Fig. 7 presents the BF taxonomy for this vulnerability.

c) The Fix: To fix the bug, the Rust community opted to set the `SmallVec` length to `index`, call `insert_many()`, and then update the length. With this fix, if an iterator panics, a memory leak occurs [12]. The developers downgraded the bug to avoid double free as a consequence, which could lead to arbitrary code execution. Now they have a memory leak. Fig. 8 presents the BF taxonomy for the new bug.



Operation			Object	
Mechanism: • Explicit	Source Code: • Standard Library (lib.rs)	Execution Space: • Userland	Ownership: • Shared	Location: • Heap

Fig. 7: BF for CVE-2018-20991 – Rust Iterator Panic



Operation			Object	
Mechanism: • Explicit	Source Code: • Standard Library (lib.rs)	Execution Space: • Userland	Ownership: • Shared	Location: • Heap

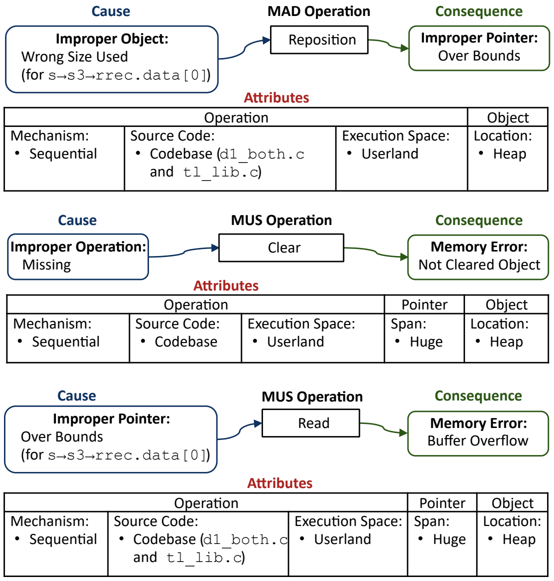
Fig. 8: BF for the Bug in the Fix of CVE-2018-20991

B. CVE-2014-0160 – Heartbleed Buffer Overflow

This vulnerability is listed in CVE-2014-0160 and discussed in [13]. The source code could be found at [14].

a) Brief Description: Heartbleed is a vulnerability due to a bug in the OpenSSL – a crypto library for the Transport Layer Security (TLS) and Secure Sockets Layer (SSL) protocols. Using the heartbeat extension tests in TLS and Datagram Transport Layer Security (DTLS) protocols, a user can send a heartbeat request to a server. The request contains a string and a `payload` unsigned integer, which value is expected to be the string size. The server responds with the same string. However, due to the bug, a malicious user could set the `payload` as big as `65535` and the server would read out of bounds. This could expose confidential information that was not cleared before release.

b) Analysis: The TLS and DTLS implementations in OpenSSL 1.0.1 before 1.0.1g have a bug in the `d1_both.c` and `t1_lib.c` files. In the Heartbleed attack, the software stores the user data in an array `s->s3->rrec.data[0]`. The size of that array is much less than the huge `65535` bytes payload. The software does not check the size of the data (`s->s3->rrec.length`) towards the value of the payload. It assumes these numbers are equal and using `memcpy()` reads `payload` consecutive bytes from the array, beginning at its first byte, then sends them to the malicious user. Fig. 9 presents the BF taxonomy for this vulnerability.



Operation			Object
Mechanism: • Sequential	Source Code: • Codebase (d1_both.c and t1_lib.c)	Execution Space: • Userland	Location: • Heap

Operation			Pointer	Object
Mechanism: • Sequential	Source Code: • Codebase	Execution Space: • Userland	Span: • Huge	Location: • Heap

Operation			Pointer	Object
Mechanism: • Sequential	Source Code: • Codebase (d1_both.c and t1_lib.c)	Execution Space: • Userland	Span: • Huge	Location: • Heap

Fig. 9: BF for CVE-2014-0160 – Heartbleed Buffer Overflow

c) The Fix: To fix the bug the openSSL team added a bound check for the array size [15]. We should note that in Fig. 9 the Wrong Size Used cause is a consequence from a missing Verify operation of a preceding Data Verification Bug (DVR) [1], which is beyond the scope of this paper.

C. Discussion

The BF taxonomy of a vulnerability can help identify exploit mitigation techniques for a particular weakness types. For that we should connect the BF taxonomy to an appropriate attack model.

For memory bugs, we can use the classic memory corruption attack model of Szekeres et al. [16] that systematizes the memory protection techniques. The model has six steps towards the ultimate goal of an attacker. Its very first level is on memory safety, where an attacker can start an exploitation with an invalid pointer dereference. This kind of invalid pointer corresponds to the improper pointer states that define some of the causes for the BF Memory Bugs classes (Table IVa).

Using the BF description of a vulnerability and following the attack model we can identify effective mitigations against possible attacks. Let's take, for example, a Buffer Overflow that is caused by Read Over Bounds. Following the Szekeres et al. model, such a bug would allow an attacker to access program data, leading to information leakage.

To make use of the collected data, the attacker should be able to interpret it. Probabilistic methods such as data space randomization (DSR) could mitigate the attack, while an address space location randomization (ASLR) will not do it [17]. The values of the Location and Execution Space attributes of the object help identify where in the memory layout the mitigation technique should be put in place.

The Szekeres et al. model, however does not cover bugs related to some BF memory operations, such as allocation, reallocation, and initialization. It does not cover any memory addressing bugs (MAD) and it is not concerned describing how a pointer becomes invalid. A key point here is that Szekeres et al. look at memory corruption bugs from attacks perspective, while we focus on systematizing information that is sufficient to fix a bug.

VII. Conclusion

In this paper, we introduce four new BF classes: Memory Addressing Bugs (MAD), Memory Allocation Bugs (MAL), Memory Use Bugs (MUS), and Memory Deallocation Bugs (MDL). We present their operations, along with the possible causes, consequences, attributes, and sites.

We analyze particular vulnerabilities related to these classes and provide precise BF descriptions. The BF structured taxonomies of memory corruption vulnerabilities show the initial error (the bug) providing a quite concise and still far more clear description than the unstructured explanations in current repositories, advisories, and publications.

Linking the BF Memory Bugs model and taxonomy to an attack model (e.g. Szekeres et al. model) would provide the means of covering the memory corruption vulnerabilities landscape. For example, the first layer of the Szekeres model could connect with the BF causes defined in Section IV-B. As part of that, the notion of invalid pointer should not be restricted to dangling pointers and out of bounds pointers; refinement to the causes in Table IVa should be considered.

The BF Memory Bugs taxonomy can be used by bug reporting tools, as it can be viewed as a structured extension over the memory related CWEs [2]. Furthermore, the BF descriptions of particular vulnerabilities can be used to identify exploit mitigation techniques.

References

[1] *The Bugs Framework*, 2020. [Online]. Available: https://samate.nist.gov/BF/.

[2] MITRE, *Common weakness enumeration (CWE)*, Accessed: 2020-02-29, 2020. [Online]. Available: https://cwe.mitre.org.

[3] ——, *Common vulnerabilities and exposures (CVE)*, Accessed: 2020-02-29, 2020. [Online]. Available: https://cve.mitre.org/.

[4] NVD, *National vulnerability database (NVD)*, Accessed: 2020-01-10, 2020. [Online]. Available: https://nvd.nist.gov.

[5] J. Caballero, G. Grieco, M. Marron, and A. Nappa, "Undangle: Early detection of dangling pointers in use-after-free and double-free vulnerabilities," in *Proc. of ISSTA*, ACM, 2012, pp. 133–143.

[6] F. Heckenbach, Accessed: 2020-02-29, 2015. [Online]. Available: https://gcc.gnu.org/bugzilla/show_bug.cgi?id=66139.

[7] V. Vikas, *Memory layout of C for beginners*, Accessed: 2021-02-18, 2019. [Online]. Available: https://medium.com/@vikasv210/memory-layout-in-c-fe4dffdaeed6.

[8] *RUSTSEC-2018-0003*, Accessed: 2020-02-29, 2018. [Online]. Available: https://rustsec.org/advisories/RUSTSEC-2018-0003.html.

[9] *rust-smallvec - Panic-safety fixes #103*, Accessed: 2020-02-29, 2014. [Online]. Available: https://github.com/servo/rust-smallvec/pull/103/files.

[10] The Rust Team, *The rustonomicon*, Accessed: 2020-04-06, 2020. [Online]. Available: https://doc.rust-lang.org/nomicon/unwinding.html.

[11] G. Hoare, *Rust*, Accessed: 2020-02-29, 2020. [Online]. Available: https://www.rust-lang.org/.

[12] *SmallVec::insert_many is unsound #96*, Accessed: 2020-02-29, 2018. [Online]. Available: https://github.com/servo/rust-smallvec/issues/96.

[13] *The heartbleed bug*, Accessed: 2020-02-29, 2014. [Online]. Available: https://heartbleed.com/.

[14] *OpenSSL*, Accessed: 2020-02-29, 2014. [Online]. Available: https://git.openssl.org/gitweb/?p=openssl.git;a=blob;f=ssl/d1_both.c;h=0a84f957118afa9804451add380eca4719a9765e;hb=4817504d069b4c5082161b02a22116ad75f822b1.

[15] *OpenSSL*, Accessed: 2020-02-29, 2014. [Online]. Available: https://github.com/openssl/openssl/commit/96db9023b881d7cd9f379b0c154650d6c108e9a3.

[16] L. Szekeres, M. Payer, T. Wei, and D. Song, "Sok: Eternal war in memory," in *Proc. of 2013 IEEE Symposium on Security and Privacy*, 2013, pp. 48–62.

[17] D. Kuvaiskii, O. Oleksenko, S. Arnautov, B. Trach, P. Bhatotia, P. Felber, and C. Fetzer, "Sgxbounds: Memory safety for shielded execution," in *Proc. of 12th European Conf. on Computer Systems*, ACM, 2017.

Bojanova, Irena. "Classifying Memory Bugs Using Bugs Framework Approach." Presented at 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC 2021). July 12, 2021 - July 16, 2021.

Improvements of Algebraic Attacks for solving the Rank Decoding and MinRank problems

Magali Bardet[4,5], Maxime Bros[1], Daniel Cabarcas[6], Philippe Gaborit[1], Ray Perlner[2], Daniel Smith-Tone[2,3], Jean-Pierre Tillich[4], and Javier Verbel[6]

[1] Univ. Limoges, CNRS, XLIM, UMR 7252, F-87000 Limoges, France
maxime.bros@unilim.fr
[2] National Institute of Standards and Technology, USA
[3] University of Louisville, USA
[4] Inria, 2 rue Simone Iff, 75012 Paris, France
[5] LITIS, University of Rouen Normandie, France
[6] Universidad Nacional de Colombia Sede Medellín, Medellín, Colombia

Abstract. Rank Decoding is the main underlying problem in rank-based cryptography. Based on this problem and quasi-cyclic versions of it, very efficient schemes have been proposed recently, such as those in the ROLLO and RQC submissions, which have reached the second round of the NIST Post-Quantum Cryptography Standardization Process. Two main approaches have been studied to solve the Rank Decoding problem: combinatorial ones and algebraic ones. While the former has been studied extensively in [24] and [10], a better understanding of the latter was recently obtained with [11] where it appeared that algebraic attacks can often be more efficient than combinatorial ones for cryptographic parameters. In particular, the results of [11] were based on Gröbner basis computations which led to complexity bounds slightly smaller than the claimed security of ROLLO and RQC cryptosystems. This paper gives substantial improvements upon this attack together with a much more precise analysis of its complexity compared to the one in [11]. Against ROLLO-I-128, ROLLO-I-192, and ROLLO-I-256, our attack has bit complexity respectively in 71, 87, and 151, to be compared to 117, 144, and 197 for the attack in [11]. Moreover, unlike this previous attack, ours does not need generic Gröbner basis algorithms since it only requires to solve a linear system. This improvement relies upon a modeling slightly different from the one in [11] combined with a new modeling for a generic MinRank instance. The latter modeling allows us to solve the MinRank problem using only linear algebra as well and no longer generic Gröbner basis algorithms, in addition to this, this new algorithm enables us to refine the analysis of MinRank's complexity given in [38]. MinRank is a problem of great interest for all multivariate-based cryptosystems, including GeMSS and Rainbow, which are at the second round of the aforementioned NIST competition; our new approach supersedes previous attacks for the MinRank problem.

Keywords: Post-quantum cryptography · NIST-PQC candidates · rank metric code-based cryptography · algebraic attack.

Bardet, Magali; Bros, Maxime; Cabarcas, Daniel; Gaborit, Philippe; Perlner, Ray; Smith-Tone, Daniel; Tillich, Jean-Pierre; Verbel, Javier.
"Improvements of Algebraic Attacks for solving the Rank Decoding and MinRank problems." Presented at 26th Annual International Conference on the Theory and Application of Cryptology and Information Security (Asiacrypt 2020). December 07, 2020 - December 11, 2020.

1 Introduction

Rank metric code-based cryptography. In the last decade, rank metric code-based cryptography has proved to be a powerful alternative to more traditional code-based cryptography based on the Hamming metric. This thread of research started with the GPT cryptosystem [22] based on Gabidulin codes [21], which are rank metric analogues of Reed-Solomon codes. However, the strong algebraic structure of those codes was successfully exploited for attacking the original GPT cryptosystem and its variants with the Overbeck attack [34] (see for example [32] for one of the latest related developments). This has to be traced back to the algebraic structure of Gabidulin codes that makes masking extremely difficult; one can draw a parallel with the situation in the Hamming metric where essentially all McEliece cryptosystems based on Reed-Solomon codes or variants of them have been broken. However, recently a rank metric analogue of the NTRU cryptosystem from [28] has been designed and studied, starting with the pioneering paper [23]. Roughly speaking, the NTRU cryptosystem relies on a lattice that has vectors of rather small Euclidean norm. It is precisely those vectors that allow an efficient decoding/deciphering process. The decryption of the cryptosystem proposed in [23] relies on LRPC codes that have rather short vectors in the dual code, but this time for the rank metric. These vectors are used for decoding in the rank metric. This cryptosystem can also be viewed as the rank metric analogue of the MDPC cryptosystem [31] that relies on short vectors in the dual code for the Hamming metric.

This new way of building rank metric code-based cryptosystems has led to a sequence of proposals [23,25,5,6], culminating in submissions to the National Institute of Standards and Technology (NIST) post-quantum competition [2,3], whose security relies solely on the decoding problem in rank metric codes with a ring structure similar to the ones encountered right now in lattice-based cryptography. Interestingly enough, one can also build signature schemes using the rank metric; even though early attempts which relied on masking the structure of a code [26,9] have been broken [16], a promising recent approach [8] only considers random matrices without structural masking.

Decoding $\mathbb{F}_{q^m}$-linear codes in Rank metric. In other words, in rank metric code-based cryptography we are now only left with assessing the difficulty of the decoding problem for the rank metric. The trend in rank metric code-based cryptography has been to consider a particular form of codes that are linear codes of length n over an extension $\mathbb{F}_{q^m}$ of degree m of $\mathbb{F}_q$, that is, $\mathbb{F}_{q^m}$-linear subspaces of $\mathbb{F}_{q^m}^n$. Let $(\beta_1, \dots, \beta_m)$ be any basis of $\mathbb{F}_{q^m}$ as a $\mathbb{F}_q$-vector space. Then words of those codes can be interpreted as matrices with entries in the ground field $\mathbb{F}_q$ by viewing a vector $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{F}_{q^m}^n$ as a matrix $\text{Mat}(\boldsymbol{x}) = (X_{ij})_{i,j}$ in $\mathbb{F}_q^{m \times n}$, where $(X_{ij})_{1 \leq i \leq m}$ is the column vector formed by the coordinates of x_j in the basis $(\beta_1, \dots, \beta_m)$, that is, $x_j = \beta_1 X_{1j} + \cdots + \beta_m X_{mj}$.

Then the “rank” metric d on $\mathbb{F}_{q^m}^n$ is the rank metric on the associated matrix space, namely

$$d(\boldsymbol{x},\boldsymbol{y}) := |\boldsymbol{y}-\boldsymbol{x}|, \quad \text{where we define } |\boldsymbol{x}| := \text{Rank}\left(\text{Mat}(\boldsymbol{x})\right).$$

Hereafter, we will use the following terminology.

Problem 1 ((m,n,k,r)-decoding problem).
Input: an $\mathbb{F}_{q^m}$-basis $(\boldsymbol{c}_1,\dots,\boldsymbol{c}_k)$ of a subspace $\mathcal{C}$ of $\mathbb{F}_{q^m}^n$, an integer $r \in \mathbb{N}$, and a vector $\boldsymbol{y} \in \mathbb{F}_{q^m}^n$ at distance at most r of $\mathcal{C}$ (i.e. $|\boldsymbol{y}-\boldsymbol{c}| \leq r$ for some $\boldsymbol{c} \in \mathcal{C}$).
Output: $\boldsymbol{c} \in \mathcal{C}$ and $\boldsymbol{e} \in \mathbb{F}_{q^m}^n$ such that $\boldsymbol{y} = \boldsymbol{c}+\boldsymbol{e}$ and $|\boldsymbol{e}| \leq r$.

This problem is known as the Rank Decoding problem, written RD. It is equivalent to the Rank Syndrome Decoding problem, written RSD, for which one uses the parity check matrix of the code. There are two approaches to solve RD instances: the combinatorial ones such as those in [24] and [10] and the algebraic ones, such as in [11]. For some time it was thought that the combinatorial approach was the most threatening attack on such schemes especially when q is small and all the parameters rank-metric submissions, until it became apparent in [11] that even for $q = 2$ the algebraic attacks outperform the combinatorial ones. Roughly speaking, if the conjecture made in [11] holds, the complexity of solving by algebraic attacks the decoding problem is of order $2^{O(r \log n)}$ with a constant depending on the rate $R = k/n$ of the code.

Even if the decoding problem is not known to be NP-complete for these $\mathbb{F}_{q^m}$-linear codes, there is a randomized reduction to an NP-complete problem [27] (namely to decoding in the Hamming metric). The region of parameters which is of interest for the NIST submissions corresponds to $m = \Theta\left(n\right)$, $k = \Theta\left(n\right)$ and $r = \Theta\left(\sqrt{n}\right)$.

The MinRank problem. The MinRank problem was first mentioned in [13] where its NP-completeness was also proven. We will consider here the homogeneous version of this problem which corresponds to

Problem 2 (MinRank problem).
Input: an integer $r \in \mathbb{N}$ and K matrices $\boldsymbol{M}_1,\dots,\boldsymbol{M}_K \in \mathbb{F}_q^{m\times n}$.
Output: field elements $x_1,x_2,\dots,x_K \in \mathbb{F}_q$ that are not all zero such that

$$\text{Rank}\left(\sum_{i=1}^{K} x_i \boldsymbol{M}_i\right) \leq r.$$

It plays a central role in public key cryptography. Many multivariate schemes are either directly based on the hardness of this problem [15] or strongly related to it as in [35,37,36] or as in the NIST post-quantum competition candidates Gui [17], GeMSS [14] or Rainbow [18]. It first appeared in this context as part of an attack against the HFE cryptosystem [35] by Kipnis and Shamir[29]. It is also central in rank metric code-based cryptography, because the RD problem reduces to MinRank as explained in [19] and actually the best algorithms for solving this problem are really MinRank solvers taking advantage of the $\mathbb{F}_{q^m}$ underlying structure as in [11]. However the parameter region generally differs. When the RD problem arising from rank metric schemes is treated as a MinRank problem we generally have $K = \theta(n^2)$ and r is rather small (often $r = \theta(\sqrt{n})$) whereas for the multivariate cryptosystems $K = \theta(n)$ but r is much bigger.

The current best known algorithms for solving the MinRank problem have exponential complexity. Many of them are obtained by an algebraic approach too consisting in modeling the MinRank problem by a system of multivariate polynomial equations and then solve it with Gröbner basis techniques. The main modelings are the Kipnis-Shamir modeling [29] and the minors modeling [20]. The complexity of solving MinRank using these modelings has been investigated in [19,20,38]. In particular [38] shows that the bilinear system that arises from the Kipnis-Shamir modeling behaves much better than generic bilinear systems with respect to Gröbner basis techniques.

Our contribution. In this paper, we follow on from the approach in [11] and propose a slightly different modeling to solve the RD problem. Roughly speaking the algebraic approach followed by [11] is to set up a bilinear system which is satisfied by the error we are looking for. This system is formed by two kinds of variables, the "coefficient" variables and the "support" variables which is implicitly the modeling considered in [33]. The breakthrough obtained in [11] was to realize that

- the coefficient variables have to satisfy "maximal minor" equations: the maximal minors of a certain $r \times (n-k-1)$ matrix (i.e. the $r \times r$ minors) with entries being linear forms in the coefficient variables have to be equal to 0.
- these maximal minors are themselves linear combinations of maximal minors c_T of an $r \times n$ matrix $\boldsymbol{C}$ whose entries are the coefficient variables.

This gives a linear system involving the c_T's and allows to find the c_T's provided that there are enough linear equations. Moreover the original bilinear system has many solutions and there is some freedom of choosing the coefficient variables and the support variables. With the choice made in [11] the information we obtain in this way about the minors of $\boldsymbol{C}$ is not enough to be able to recover the coefficient variables directly (i.e. the entries of $\boldsymbol{C}$). In this case the last step of the algebraic attack still has to compute a Gröbner basis for the algebraic system consisting of the original system plus the knowledge on the c_T's obtained from the linear system.

The new approach followed in this paper uses the fact that there is a better way to use the freedom on the coefficient variables and the support variables: we can actually specify so many coefficient variables that all the remaining entries that we do not know are essentially equal to some maximal minor c_T of $\boldsymbol{C}$. This approach allows to avoid completely the computation of the Gröbner basis: we obtain from the knowledge of the c_T's obtained from the aforementioned linear system the coefficient variables and plugging in theses values in the original bilinear system it just remains to solve a linear system involving the support variables. This new approach brings on a substantial speed-up in the computations for solving the system. It results in the best practical efficiency and complexity bounds that are currently known for the decoding problem; in particular, it significantly improves upon the aforementioned similar approach in [11]. We present attacks for ROLLO-I-128, ROLLO-I-192, and ROLLO-I-256 with bit complexity respectively in 70, 86, and 158, to be compared to 117, 144, and 197 for the attack in [11]. The difference with [11] is significant since as there is no real quantum speed-up for solving linear systems, the best quantum attacks for ROLLO-I-192 remained the quantum attack based on combinatorial attacks, when our new attacks show that ROLLO parameters are broken and need to be changed.

Our analysis is divided into two categories: the "overdetermined" and the "underdetermined" case. An (m, n, k, r)-decoding instance is overdetermined if the condition

$$m\binom{n-k-1}{r} \geq \binom{n}{r} - 1 \tag{1}$$

is fulfilled. This really corresponds to the case where we have enough linear equations by our approach to find all the c_T's (and hence all the coefficient variables). In that case we obtain a complexity in

$$\mathcal{O}\left(m\binom{n-p-k-1}{r}\binom{n-p}{r}^{\omega-1}\right) \tag{2}$$

operations in the field $\mathbb{F}_q$, where ω is the constant of linear algebra and $p = \max\{i : i \in \{1..n\}, m\binom{n-i-k-1}{r} \geq \binom{n-i}{r} - 1\}$ represents, in case the overdetermined condition (1) is comfortably fulfilled, the use of punctured codes. This complexity clearly supersedes the previous results of [11] in terms of complexity and also by the fact that it does not require generic Gröbner Basis algorithms. In a rough way for $r = \mathcal{O}(\sqrt{n})$ (the type of parameters used for ROLLO and RQC), the recent improvements on algebraic attacks can be seen as this: before [11] the complexity for solving RD involved a term in $O(n^2)$ in the upper part of a binomial coefficient, the modeling in [11] replaced it by a term in $\mathcal{O}\left(n^{\frac{3}{2}}\right)$ whereas our new modeling involves a term in $\mathcal{O}(n)$ at a similar position. This leads to a gain in the exponential coefficient of order 30 % compared to [11] and of order 50 % compared to approaches before [11]. Notice that for ROLLO and RQC only parameters with announced complexities 128 and 192 bits satisfied condition (1) but not parameters with announced complexities 256 bits.

When condition (1) is not fulfilled, the instance can either be underdetermined or be brought back to the overdetermined area by an hybrid approach using exhaustive search with exponential complexity to guess few variables in the system. In the underdetermined case, our approach is different from [11]. Here we propose an approach using reduction to the MinRank problem and a new way to solve it. Roughly speaking we start with a quadratic modeling of MinRank that we call "support minors modeling" which is bilinear in the aforementioned coefficient and support variables and linear in the so called "linear variables". The last ones are precisely the x_i's that appear in the MinRank problem. Recall that the coefficient variables are the entries of a $r \times n$ matrix $\boldsymbol{C}$. The crucial observation is now that for all positive integer b all maximal minors of any $(r+b) \times n$ matrix obtained by adding to $\boldsymbol{C}$ any b rows of $\sum_i x_i \boldsymbol{M}_i$ are equal to 0. These minors are themselves linear combinations of terms of the form mc_T where c_T is a maximal minor of $\boldsymbol{C}$ and m a monomial of degree b in the x_i's. We can predict the number of independent linear equations in the mc_T's we obtain this way and when the number of such equations is bigger than the number of mc_T's we can recover their values and solve the MinRank problem without computing Gröbner bases. This new approach is not only effective in the underdetermined case of the RD problem it can also be quite effective for some multivariate proposals made to the NIST competition. In the case of the RD problem, it improves the attacks on [7] made in [11] for the parameter sets with the largest values of r (corresponding to parameters claiming 256 bits of security). The multivariate schemes that are affected by this new attack are for instance GeMSS and Rainbow. On GeMSS it shows MinRank attacks together with this new way of solving MinRank come close to the best known attacks against this scheme. On Rainbow it outperforms slightly the best known attacks for certain high security parameter sets.

At last, not only do these two new ways of solving algebraically the RD or MinRank problem outperform previous algebraic approaches in certain parameter regimes, they are also much better understood: we do not rely on heuristics based on the the first degree fall as in [38,11] to analyze its complexity, but it really amounts to solve a linear system and understand the number of independent linear equations that we obtain which is something for which we have been able to give accurate formulas predicting the behavior we obtain experimentally.

2 Notation

In what follows, we use the following notation and definitions:

- Matrices and vectors are written in boldface font $\boldsymbol{M}$.
- The transpose of a matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{M}^\intercal$.
- For a given ring $\mathcal{R}$, the set of matrices with n rows, m columns and coefficients in $\mathcal{R}$ is denoted by $\mathcal{R}^{n \times m}$.
- $\{1..n\}$ stands for the set of integers from 1 to n.
- For a subset $I \subset \{1..n\}$, $\#I$ stands for the number of elements in I.
- For two subsets $I \subset \{1..n\}$ and $J \subset \{1..m\}$, we write $\boldsymbol{M}_{I,J}$ for the submatrix of $\boldsymbol{M}$ formed by its rows (resp. columns) with index in I (resp. J).
- We use the shorthand notations $\boldsymbol{M}_{*,J} = \boldsymbol{M}_{\{1..m\},J}$ and $\boldsymbol{M}_{I,*} = \boldsymbol{M}_{I,\{1..n\}}$, where $\boldsymbol{M}$ has m rows and n columns, and $\boldsymbol{M}_{i,j}$ for the entry in row i and column j.
- We denote the determinant of a matrix $\boldsymbol{M}$ by $|\boldsymbol{M}|$. We also use a notation inspired by the previous one for denoting the determinant of a submatrix, $|\boldsymbol{M}|_{I,J}$ denotes the determinant of the submatrix $\boldsymbol{M}_{I,J}$ and $|\boldsymbol{M}|_{*,J}$ denotes the principal minor of $\boldsymbol{M}$ obtained by taking the determinant of $\boldsymbol{M}_{*,J}$.

- $\alpha \in \mathbb{F}_{q^m}$ is a primitive element, that is to say that $(1, \alpha, \dots, \alpha^{m-1})$ is a basis of $\mathbb{F}_{q^m}$ seen as an $\mathbb{F}_q$-vector space.
- For $\boldsymbol{v} = (v_1, \dots, v_n) \in \mathbb{F}_{q^m}^n$, the *support* of $\boldsymbol{v}$ is the $\mathbb{F}_q$-vector subspace of $\mathbb{F}_{q^m}$ spanned by the vectors $v_1, \dots, v_n$. Thus this support is the column space of the matrix $\mathrm{Mat}(\boldsymbol{v})$ associated to $\boldsymbol{v}$ (for any choice of basis), and its dimension is precisely $\mathrm{Rank}(\mathrm{Mat}(\boldsymbol{v}))$.
- An $[n, k]$ $\mathbb{F}_{q^m}$-linear code is an $\mathbb{F}_{q^m}$-linear subspace of $\mathbb{F}_{q^m}^n$ of dimension k.
- Unless otherwise specified, the *decoding problem* always refers to the Rank Decoding problem.

3 Algebraic modeling of the MinRank and the decoding problem

3.1 Modeling of MinRank

Here is the modeling for the MinRank problem that we consider, it is related to the modeling used for decoding in the rank metric in [11]. The starting point is that, in order to solve the equation in *Problem* 2, we look for a nonzero solution $(\boldsymbol{S}, \boldsymbol{C}, \boldsymbol{x}) \in \mathbb{F}_q^{m \times r} \times \mathbb{F}_q^{r \times n} \times \mathbb{F}_q^K$ of

$$\boldsymbol{S}\boldsymbol{C} = \sum_{i=1}^{K} x_i \boldsymbol{M}_i. \tag{3}$$

$\boldsymbol{S}$ is an unknown matrix whose columns give a basis for the column space of a matrix of rank $\leq r$ we are looking for (i.e. $\sum_{i=1}^{K} x_i \boldsymbol{M}_i$). The i-th column of $\boldsymbol{C}$ represents the coordinates of the i-th column of the aforementioned matrix in this basis. We call the entries of $\boldsymbol{S}$ the *support variables*, and the entries of $\boldsymbol{C}$ the *coefficient variables*. Note that in the above equation, the variables x_i only occur linearly. As such, we will dub them the *linear variables*.

Let $\boldsymbol{r}_j$ be the j-th row of the matrix $\sum_{i=1}^{K} x_i \boldsymbol{M}_i$. Equation (3) implies that each row $\boldsymbol{r}_j$ is in the rowspace of $\boldsymbol{C}$ (or in coding theoretic terms $\boldsymbol{r}_j$ should belong to the code $\mathcal{C}$ generated by $\boldsymbol{C}$, that is $\mathcal{C} := \{\boldsymbol{u}\boldsymbol{C}, \boldsymbol{u} \in \mathbb{F}_q^r\}$). This implies that the following $(r+1) \times n$ matrix $\boldsymbol{C}'_j$ is of rank $\leq r$:

$$\boldsymbol{C}'_j = \begin{pmatrix} \boldsymbol{r_j} \\ \boldsymbol{C} \end{pmatrix}.$$

Therefore, all the maximal minors of this matrix should be equal to 0. Notice that these maximal minors can be expressed via cofactor expansion with respect to their first row. In this way, they can be seen as bilinear forms in the variables x_i and the $r \times r$ minors of $\boldsymbol{C}$. These minors will play a fundamental role in the whole paper and we will use the following notation for them.

Notation 1 *Let $T \subset \{1..n\}$ with $\#T = r$. We let*

$$c_T := |\boldsymbol{C}|_{*,T}$$

be the maximal minor of $\boldsymbol{C}$ corresponding to the columns of $\boldsymbol{C}$ that belong to T.

These considerations lead to the following algebraic modeling.

Modeling 1 (Support Minors modeling) *We consider the system of bilinear equations, given by canceling the maximal minors of the* m *matrices* $\boldsymbol{C}'_j$*:*

$$\left\{ f = 0 \middle| f \in \mathbf{MaxMinors} \begin{pmatrix} \boldsymbol{r_j} \\ \boldsymbol{C} \end{pmatrix}, \ j \in \{1..m\} \right\}. \tag{4}$$

This system contains:

- $m\binom{n}{r+1}$ *bilinear equations with coefficients in* $\mathbb{F}_q$*,*
- $K + \binom{n}{r}$ *unknowns:* $\boldsymbol{x} = (x_1, \cdots, x_K)$ *and the* c_T*'s,* $T \subset \{1..n\}$ *with* $\#T = r$*.*

We search for the solutions x_i, c_T*'s in* $\mathbb{F}_q$*.*

Remark 1.

1. One of the point of having the c_T as unknowns instead of the coefficients C_{ij} of $\boldsymbol{C}$ is that, if we solve (4) directly in the x_i and the C_{ij} variables, then there are actually plenty of solutions to (4) since when $(\boldsymbol{x}, \boldsymbol{C})$ is a solution for it, then $(\boldsymbol{x}, \boldsymbol{AC})$ is also a solution for any invertible matrix $\boldsymbol{A}$ in $\mathbb{F}_q^{r \times r}$. With the c_T variables we only expect a space of dimension 1 for the c_T corresponding to the transformation $c_T \mapsto |\boldsymbol{A}| \, c_T$ that maps a given solution of (4) to a new one.
2. Another benefit brought by replacing the C_{ij} variables by the c_T's is of course that it brings a big saving in the number of possible monomials for writing the algebraic system (4) (about $r!$ times less). This allows for instance for solving this system by direct linearization when the number of equations of the previous modeling is larger than or equal to the number of different $x_i c_T$ monomials minus 1, namely when

$$m\binom{n}{r+1} \geq K\binom{n}{r} - 1. \tag{5}$$

This turns out to be "almost" the case for several multivariate cryptosystem proposals based on the MinRank problem where K is generally of the same order as m and n.

3.2 The approach followed in [11] to solve the decoding problem

In what follows, we consider the (m, n, k, r)-decoding problem for a code $\mathcal{C}$ of length n, dimension k over $\mathbb{F}_{q^m}$ and assume we have received $\boldsymbol{y} \in \mathbb{F}_{q^m}^n$ at distance r from $\mathcal{C}$ and look for $\boldsymbol{c} \in \mathcal{C}$ and $\boldsymbol{e}$ such that $\boldsymbol{y} = \boldsymbol{c} + \boldsymbol{e}$ and $|\boldsymbol{e}| = r$. We will assume in what follows that there is a unique solution to this problem (which is relevant for our cryptographic schemes). The starting point is the Ourivksi-Johansson approach, consisting in considering the linear code $\widetilde{C} = \mathcal{C} + \langle \boldsymbol{y} \rangle$. By construction, $\boldsymbol{e}$ belongs to $\widetilde{C}$ as well as all its multiples $\lambda \boldsymbol{e}$, $\lambda \in \mathbb{F}_{q^m}$. Looking for non-zero codewords in $\widetilde{C}$ of rank weight r has at least $q^m - 1$ different solutions, namely all the $\lambda \boldsymbol{e}$ for $\lambda \in \mathbb{F}_{q^m}^{\times}$.

It is readily seen that finding such codewords can be done by solving the (homogeneous) MinRank problem with $\boldsymbol{M}_{ij} := \mathrm{Mat}(\alpha^{i-1}\boldsymbol{c}_j)$ (we adopt a bivariate indexing of the $\boldsymbol{M}_i$'s which is more convenient here), for $(ij) \in \{1..m\}\times\{1..k+1\}$ and where $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{k+1}$ is an $\mathbb{F}_{q^m}$-basis of $\widetilde{C}$. This is a consequence of the fact that the $\alpha^{i-1}\boldsymbol{c}_j$'s form an $\mathbb{F}_q$-basis of $\widetilde{C}$. However, the problem with this approach is that $K = (k+1)m$ which is of order $\Omega(n^2)$ for the parameters relevant to cryptography. This is much more than for the multivariate cryptosystems based on MinRank and (5) is far from being satisfied here. However, as observed in [11], it turns out in this particular case, it is possible because of the $\mathbb{F}_{q^m}$ linear structure of the code, to give an algebraic modeling that only involves the coefficients variables, that is the entries of $\boldsymbol{C}$. It is obtained by introducing a parity-check matrix for $\widetilde{C}$, that is a matrix $\boldsymbol{H}$ whose kernel is $\widetilde{C}$:

$$\widetilde{C} = \{\boldsymbol{c} \in \mathbb{F}_{q^m}^n : \boldsymbol{c}\boldsymbol{H}^\intercal = 0\}.$$

In our $\mathbb{F}_{q^m}$ linear setting the solution $\boldsymbol{e}$ we are looking for can be written as

$$\boldsymbol{e} = \begin{pmatrix}1 \; \alpha \; \dots \; \alpha^{m-1}\end{pmatrix} \boldsymbol{S}\boldsymbol{C}, \tag{6}$$

where $\boldsymbol{S} \in \mathbb{F}_q^{m\times r}$ and $\boldsymbol{C} \in \mathbb{F}_q^{r\times n}$ play the same role as in the previous subsection: $\boldsymbol{S}$ represents a basis of the support of $\boldsymbol{e}$ in $\left(\mathbb{F}_q^m\right)^r$ and $\boldsymbol{C}$ the coordinates of $\boldsymbol{e}$ in this basis. By writing that $\boldsymbol{e}$ should belong to $\widetilde{C}$ we obtain that

$$\begin{pmatrix}1 \; \alpha \; \dots \; \alpha^{m-1}\end{pmatrix} \boldsymbol{S}\boldsymbol{C}\boldsymbol{H}^\intercal = \boldsymbol{0}_{n-k-1}. \tag{7}$$

The algebraic system involving only the coefficient variables follows immediately from this.

Proposition 1 ([11], Theorem 2). *The maximal minors of the* $r \times (n-k-1)$ *matrix* $\boldsymbol{C}\boldsymbol{H}^\intercal$ *are all equal to* 0.

Proof. Consider the following vector in $\mathbb{F}_q^r$: $\boldsymbol{e}' := \begin{pmatrix}1 \; \alpha \; \dots \; \alpha^{m-1}\end{pmatrix} \boldsymbol{S}$ whose entries generate (over $\mathbb{F}_q$) the subspace generated by the entries of $\boldsymbol{e}$ (i.e. its support). Substituting $\begin{pmatrix}1 \; \alpha \; \dots \; \alpha^{m-1}\end{pmatrix} \boldsymbol{S}$ for $\boldsymbol{e}'$ in (7) yields

$$\boldsymbol{e}'\boldsymbol{C}\boldsymbol{H}^\intercal = \boldsymbol{0}_{n-k-1}.$$

This shows that the r rows of $\boldsymbol{C}\boldsymbol{H}^\intercal$ are not independent and that the $r \times n$ matrix $\boldsymbol{C}\boldsymbol{H}^\intercal$ is of rank $\leq r-1$. □

These minors $\boldsymbol{C}\boldsymbol{H}^\intercal$ are polynomials in the entries of $\boldsymbol{C}$ with coefficients in $\mathbb{F}_{q^m}$. Since these entries belong to $\mathbb{F}_q$, the nullity of each minor gives m algebraic equations corresponding to polynomials with coefficients in $\mathbb{F}_q$. This involves the following operation

Notation 2 *Let* $\mathcal{S} := \{\sum_j a_{ij}m_{ij} = 0, 1 \leq i \leq N\}$ *be a set of polynomial equations where the* m_{ij}*'s are the monomials in the unknowns that are assumed to belong to* $\mathbb{F}_q$*, whereas the* a_{ij}*'s are known coefficients that belong to* $\mathbb{F}_{q^m}$*. We* *define the a_{ijk}'s as $a_{ij} = \sum_{k=0}^{m-1} a_{ijk}\alpha^k$, where the a_{ijk}'s belong to $\mathbb{F}_q$. From this we can define the system "unfolding" over $\mathbb{F}_q$ as*

$$\mathbf{UnFold}\,(\mathcal{S}) := \left\{ \sum_j a_{ijk} m_{ij} = 0, 1 \leq i \leq N, 0 \leq k \leq m-1 \right\}.$$

The important point is that the solutions of $\mathcal{S}$ over $\mathbb{F}_q$ are exactly the solutions of $\mathbf{UnFold}\,(\mathcal{S})$ over $\mathbb{F}_q$, so that in that sense the two systems are equivalent.

By using the Cauchy-Binet formula, it is proved [11, Prop. 1] that the maximal minors of $\boldsymbol{C}\boldsymbol{H}^{\intercal}$, which are polynomials of degree $\leq r$ in the coefficient variables C_{ij}, can actually be expressed as *linear* combinations of the c_T's. In other words we obtain $m\binom{n-k-1}{r}$ linear equations over $\mathbb{F}_q$ by "unfolding" the $\binom{n-k-1}{r}$ maximal minors of $\boldsymbol{C}\boldsymbol{H}^{\intercal}$. We denote such a system by

$$\mathbf{UnFold}\,(\mathbf{MaxMinors}(\boldsymbol{C}\boldsymbol{H}^{\intercal}))\,. \tag{8}$$

It is straightforward to check that some variables in $\boldsymbol{C}$ and $\boldsymbol{S}$ can be specialized. The choice which is made in [11] is to specialize $\boldsymbol{S}$ with its r first rows equal to the identity ($\boldsymbol{S}_{\{1..r\},*} = \boldsymbol{I}_r$), its first column to $\mathbf{1}^{\intercal} = (1, 0, \ldots, 0)^{\intercal}$ and $\boldsymbol{C}$ has its first column equal to $\mathbf{1}^{\intercal}$. It is proved in [11, Section 3.3] that if the first coordinate of $\boldsymbol{e}$ is nonzero and the top $r \times r$ block of $\boldsymbol{S}$ is invertible, then the solution of the previous specialized system is also a solution of the system without specialization. Moreover, this will always be the case up to a permutation of the coordinates of the codewords or a change of $\mathbb{F}_{q^m}$-basis.

It is proved in [11, Prop. 2] that a degree-r Gröbner basis of the unfolded polynomials **MaxMinors** can be obtained by solving the corresponding linear system in the c_T's.

However, this strategy of specialization does not reveal directly the entries of $\boldsymbol{C}$ (it only reveals the values of the c_T's). To finish the calculation it still remains to compute a Gröbner basis of the whole algebraic system as explained in [11, Step 5, §6.1]). There is a simple way to avoid this computation by specializing the variables of $\boldsymbol{C}$ in a different way. This is the new approach we will explain now.

3.3 The new approach to solve the decoding problem : specializing the identity in $\boldsymbol{C}$

As for the previous approach, we notice that if $(\boldsymbol{S}, \boldsymbol{C})$ is a solution of (7) then $(\boldsymbol{S}\boldsymbol{A}^{-1}, \boldsymbol{A}\boldsymbol{C})$ is also a solution of the same system for any invertible matrix $\boldsymbol{A}$ in $\mathbb{F}_q^{r\times r}$. Now, in the case where the first r columns of a solution $\boldsymbol{C}$ form a invertible matrix, we will still have a solution with the specialization

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{I}_r \ \boldsymbol{C}' \end{pmatrix}.$$

We can also specialize the first column of $\boldsymbol{S}$ to $\mathbf{1}^{\intercal} = \begin{pmatrix} 1 \ 0 \ldots 0 \end{pmatrix}^{\intercal}$. If the first r columns of $\boldsymbol{C}$ are not independent, it suffices as in [11, Algorithm 1] to make several different attempts of choosing r columns. The point of this specialization is that

- the corresponding c_T's are equal to the entries C_{ij} of $\boldsymbol{C}$ up to an unessential factor $(-1)^{r+i}$ whenever $T = \{1..r\} \backslash \{i\} \cup \{j\}$ for any $i \in \{1..r\}$ and $j \in \{r+1..n\}$. This follows on the spot by writing the cofactor expansion of the minor $c_T = |\boldsymbol{C}|_{*,\{1..r\}\backslash\{i\}\cup\{j\}}$. Solving the linear system in the c_T's corresponding to (8) yields now directly the coefficient variables C_{ij}. This avoids the subsequent Gröbner basis computation, since once we have $\boldsymbol{C}$ we obtain $\boldsymbol{S}$ directly by solving (7) which has become a linear system.
- it is readily shown that any solution of (8) is actually a projection on the C_{ij} variables of a solution $(\boldsymbol{S}, \boldsymbol{C})$ of the whole system (see Proposition 3). This justifies the whole approach.

In other words we are interested here in the following modeling

Modeling 2 *We consider the system of linear equations, given by unfolding all maximal minors of* $\left(\boldsymbol{I}_r\ \boldsymbol{C}'\right)\boldsymbol{H}^\intercal$*:*

$$\left\{ f = 0 \middle| f \in \textbf{UnFold}\left(\textbf{MaxMinors}\left(\left(\boldsymbol{I}_r\ \boldsymbol{C}'\right)\boldsymbol{H}^\intercal\right)\right)\right\}. \tag{9}$$

This system contains:

- $m\binom{n-k-1}{r}$ *linear equations with coefficients in* $\mathbb{F}_q$*,*
- $\binom{n}{r} - 1$ *unknowns: the* c_T*'s,* $T \subset \{1..n\}$ *with* $\#T = r$*,* $T \neq \{1..r\}$*.*

We search for the solutions c_T*'s in* $\mathbb{F}_q$*.*

Note that from the specialization, $c_{\{1..r\}} = 1$ is not an unknown.

For the reader's convenience, let us recall the specific form of these equations which is obtained by unfolding the following polynomials (see [11, Prop. 2] and its proof).

Proposition 2. *The system* $\textbf{MaxMinors}(\boldsymbol{C}\boldsymbol{H}^\intercal)$ *contains* $\binom{n-k-1}{r}$ *polynomials of degree* r *over* $\mathbb{F}_{q^m}$*, indexed by the subsets* $J \subset \{1..n-k-1\}$ *of size* r*, that are the*

$$P_J = \sum_{\substack{T_1 \subset \{1..k+1\}, T_2 \subset J, \\ \#T_1 + \#T_2 = r \\ T = T_1 \cup (T_2 + k + 1)}} (-1)^{\sigma_J(T_2)} \left|\boldsymbol{R}\right|_{T_1, J \backslash T_2} c_T, \tag{10}$$

where the sum is over all subsets $T_1 \subset \{1..k+1\}$ *and* T_2 *subset of* J*, with* $\#T_1 + \#T_2 = r$*, and* $\sigma_J(T_2)$ *is an integer depending on* T_2 *and* J*. We denote by* $T_2 + k + 1$ *the set* $\{i + k + 1 : i \in T_2\}$*.*

Let us show now that the solutions of the linear system obtained this way are projections of the solutions of the original system. For this purpose, let us bring in

– The original system (7) over $\mathbb{F}_{q^m}$ obtained with the aforementioned specialization

$$\mathcal{F}_C = \left\{ \left(1\ \alpha \cdots \alpha^{m-1}\right) \left(\mathbf{1}^\intercal\ \boldsymbol{S}'\right) \left(\boldsymbol{I}_r\ \boldsymbol{C}'\right) \boldsymbol{H}^\intercal = \mathbf{0}_{n-k-1} \right\}, \tag{11}$$

where $\mathbf{1}^\intercal = \left(1\ 0 \ldots 0\right)^\intercal$, $\boldsymbol{S} = \left(\mathbf{1}^\intercal\ \boldsymbol{S}'\right)$ and $\boldsymbol{C} = \left(\boldsymbol{I}_r\ \boldsymbol{C}'\right)$.

– The system in the coefficient variables we are interested in

$$\mathcal{F}_M = \left\{ f = 0 \middle| f \in \mathbf{MaxMinors}\left(\left(\boldsymbol{I}_r\ \boldsymbol{C}'\right) \boldsymbol{H}^\intercal\right) \right\}, \tag{12}$$

– Let $V_{\mathbb{F}_q}(\mathcal{F}_C)$ be the set of solutions of (11) with all variables in $\mathbb{F}_q$, that is

$$V_{\mathbb{F}_q}(\mathcal{F}_C) = \\ \left\{ (\boldsymbol{S}^*, \boldsymbol{C}^*) \in {\mathbb{F}_q}^{m(r-1)+r(n-r)} : \left(1\ \alpha \cdots \alpha^{m-1}\right) \left(\mathbf{1}^\intercal\ \boldsymbol{S}^*\right) \left(\boldsymbol{I}_r\ \boldsymbol{C}^*\right) \boldsymbol{H}^\intercal = \mathbf{0} \right\}. \tag{13}$$

– Let $V_{\mathbb{F}_q}(\mathcal{F}_M)$ be the set of solutions of (12) with all variables in $\mathbb{F}_q$, i.e.

$$V_{\mathbb{F}_q}(\mathcal{F}_M) = \left\{ \boldsymbol{C}^* \in {\mathbb{F}_q}^{r(n-r)} : \mathtt{Rank}_{\mathbb{F}_{q^m}}\left(\left(\boldsymbol{I}_r\ \boldsymbol{C}^*\right) \boldsymbol{H}^\intercal\right) < r \right\}.$$

With these notations at hand, we will now show that solving the decoding problem is left to solve the **MaxMinors** system, that depends only on the $\boldsymbol{C}$ variables.

Proposition 3. *If* $\boldsymbol{e}$ *can be uniquely decoded and has rank* r*, then*

$$V_{\mathbb{F}_q}(\mathcal{F}_M) = \left\{ \boldsymbol{C}^* \in \mathbb{F}_q^{r(n-r)} : \exists \boldsymbol{S}^* \in \mathbb{F}_q^{m(r-1)} \ s.t.\ (\boldsymbol{S}^*, \boldsymbol{C}^*) \in V_{\mathbb{F}_q}(\mathcal{F}_C) \right\}. \tag{14}$$

This means that the set $V_{\mathbb{F}_q}(\mathcal{F}_M)$ *is the projection of the set* $V_{\mathbb{F}_q}(\mathcal{F}_C)$ *on the last* $r(n-r)$ *coordinates.*

Proof. Let $(\boldsymbol{S}^*, \boldsymbol{C}^*) \in V_{\mathbb{F}_q}(\mathcal{F}_C)$, then the non-zero vector

$$\left(1\ S_2^* \ldots S_r^*\right) = \left(1\ \alpha \cdots \alpha^{m-1}\right) \left(\mathbf{1}^\intercal\ \boldsymbol{S}^*\right)$$

belongs to the left kernel of the matrix $\left(\boldsymbol{I}_r\ \boldsymbol{C}^*\right) \boldsymbol{H}^\intercal$. Hence this matrix has rank less than r, and $\boldsymbol{C}^* \in V_{\mathbb{F}_q}(\mathcal{F}_M)$. Reciprocally, if $\boldsymbol{C}^* \in V_{\mathbb{F}_q}(\mathcal{F}_M)$, then the matrix $\left(\boldsymbol{I}_r\ \boldsymbol{C}^*\right) \boldsymbol{H}^\intercal$ has rank less than r, hence its left kernel over $\mathbb{F}_{q^m}$ contains a non zero element $(S_1^*, \ldots, S_r^*) = (1, \alpha, \ldots, \alpha^{m-1})\boldsymbol{S}^*$ with the coefficients of $\boldsymbol{S}^*$ in $\mathbb{F}_q$. But S_1^* cannot be zero, as it would mean that $(0, S_2^*, \ldots, S_r^*) \left(\boldsymbol{I}_r\ \boldsymbol{C}^*\right)$ is an error of weight less than r solution of the decoding problem, and we assumed there is only one error of weight exactly r solution of the decoding problem. Then, $({S_1^*}^{-1}(S_2^*, \ldots, S_r^*), \boldsymbol{C}^*) \in V_{\mathbb{F}_q}(\mathcal{F}_C)$. □

4 Solving the rank decoding problem: overdetermined case

In this section, we show that, when the number of equations is sufficiently large, we can solve the system given in modeling 2 with only linear algebra computations, by linearization on the c_T's.

4.1 The overdetermined case

The linear system given in Modeling 2 is described by the following matrix **MaxMin** with rows indexed by $(J,i) : J \subset \{1..n-k-1\}, \#J = r, 0 \leq i \leq m-1$ and columns indexed by $T \subset \{1..n\}$ of size r, with the entry in row (J,i) and column T being the coefficient in α^i of the element $\pm \left|\boldsymbol{R}\right|_{T_1, J\setminus T_2} \in \mathbb{F}_{q^m}$. More precisely, we have

$$\begin{aligned} \mathbf{MaxMin}[(J,i),T] &= \begin{cases} 0 & \text{if } T_2 \not\subset J \\ [\alpha^i](-1)^{\sigma_J(T_2)}(\left|\boldsymbol{R}\right|_{T_1,J\setminus T_2}) & \text{if } T_2 \subset J, \end{cases} \qquad (15) \\ \text{with} \quad & T_1 = T \cap \{1..k+1\}, \\ \text{and} \quad & T_2 = (T \cap \{k+2..n\}) - (k+1). \end{aligned}$$

The matrix **MaxMin** can have rank $\binom{n}{r} - 1$ at most; indeed if it had a maximal rank of $\binom{n}{r}$, this would imply that all c_T's are equal to 0, which is in contradiction with the assumption $c_{\{1..r\}} = 1$.

Proposition 4. *If* **MaxMin** *has rank* $\binom{n}{r} - 1$ *(which implies that* $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$*), then the right kernel of* **MaxMin** *contains only one element* $\begin{pmatrix}\boldsymbol{c} & 1\end{pmatrix} \in \mathbb{F}_q^{\binom{n}{r}}$ *with value 1 on its component corresponding to* $c_{\{1..r\}}$*. The components* $\boldsymbol{c}$ *of this vector contain the values of the* c_T*'s,* $T \neq \{1..r\}$*. This gives in particular the values of all the variables* $C_{i,j} = (-1)^{r+i} c_{\{1..r\}\setminus\{i\}\cup\{j\}}$*.*

Proof. If **MaxMin** has rank $\binom{n}{r} - 1$, then as there is a solution to the system, a row echelon form of the matrix has the shape

$$\begin{pmatrix} \boldsymbol{I}_{\binom{n}{r}-1} & -\boldsymbol{c}^{\intercal} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

with $\boldsymbol{c}$ a vector in $\mathbb{F}_q$ of size $\binom{n}{r} - 1$: we cannot get a jump in the stair of the echelon form as it would imply that Eq. (12) has no solution. Then $\begin{pmatrix}\boldsymbol{c} & 1\end{pmatrix}$ is in the right kernel of **MaxMin**. □

It is then easy to recover the variables $\boldsymbol{S}$ from (11) by linear algebra. The following algorithm recovers the error if there is one solution to the system (11). It is shown in [11, Algorithm 1] how to deal with the other cases, and this can be easily adapted to the specialization considered in this paper.

Proposition 5. *When* $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$ *and* **MaxMin** *has rank* $\binom{n}{r} - 1$*, then Algorithm 1 recovers the error in complexity*

$$\mathcal{O}\left(m \binom{n-k-1}{r} \binom{n}{r}^{\omega - 1} \right) \qquad (16)$$

operations in the field $\mathbb{F}_q$*, where* ω *is the constant of linear algebra.*

```
Input: Code $\mathcal{C}$, vector $\boldsymbol{y}$ at distance $r$ from $\mathcal{C}$, such that $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$
       and MaxMin has rank $\binom{n}{r} - 1$
Output: The error $\boldsymbol{e}$ of weight $r$ such that $\boldsymbol{y} - \boldsymbol{e} \in \mathcal{C}$
Construct MaxMin, the $m\binom{n-k-1}{r} \times \binom{n}{r}$ matrix over $\mathbb{F}_q$ associated to the
  system MaxMinors Eq. (12) ;
Let $(\boldsymbol{c}\ 1)$ be the only such vector in the right kernel of MaxMin ;
Compute the values $\boldsymbol{C}^* = (c^*_{i,j})_{i,j}$ from $\boldsymbol{c}$;
Compute the values $(S^*_1, \dots, S^*_r) \in \mathbb{F}^r_{q^m}$ by solving the linear system

        $(S_1, \dots, S_r)\boldsymbol{C}^*\boldsymbol{H}^\intercal = 0$

and taking the unique value with $S^*_1 = 1$;
return $(1, S^*_2, \dots, S^*_r)\boldsymbol{C}^*$ ;
```

Algorithm 1: (m, n, k, r)-decoding in the overdetermined case.

Proof. To recover the error, the most consuming part is the computation of the left kernel of the matrix **MaxMin** in $\mathbb{F}_q^{m\binom{n-k-1}{r}\times\binom{n}{r}}$, in the case where $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$. This can be done by computing an echelon form of **MaxMin**, in this case the complexity is bounded by Eq. (16). □

We ran a lot of experiments with random codes $\mathcal{C}$ such that $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$, and the matrix **MaxMin** was always of rank $\binom{n}{r} - 1$. That is why we propose the following heuristic about the rank of **MaxMin**.

Heuristic 1 (Overdetermined case) *When* $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$, *with overwhelming probability, the rank of the matrix* **MaxMin** *is* $\binom{n}{r} - 1$.

Figure 1 gives the experimental results we obtained for $q = 2$, $r = 3, 4, 5$ and different values of n. We choose to keep m prime and close to $n/1.18$ to have a data set containing the parameters of the ROLLO-I cryptosystem. We choose for k the minimum between $\frac{n}{2}$ and the largest value leading to an overdetermined case. We have $k = \frac{n}{2}$ as soon as $n \geq 22$ for $r = 3$, $n \geq 36$ for $r = 4$, $n \geq 58$ for $r = 5$. The figure shows that the estimated complexity is a good upper bound for the computation's complexity. It also shows that this upper bound is not tight. Note that the experimental values are the complexity of the whole attack, including the build of the the matrix that require to compute the minors of $\boldsymbol{R}$. Hence for small values of n, it may happen that this part of the attack takes more time than the solving of the linear system. This explains why, for $r = 3$ and $n < 28$, the experimental curve is above the theoretical one.

Figure 2 shows the theoretical complexity, in the case where $n = 2k$ and m is prime and close to $n/1.18$. We take those parameters because they fit with the parameters in the cryptosystem ROLLO-I. When the parameters (m, n, k, r) do not satisfy the condition $m\binom{n-k-1}{r} \geq \binom{n}{r} - 1$, we do not give the complexity. The graph starts from the first value of n where $(n/1.18, n, 2k, r)$ is in the overdetermined case. We can see that theoretically, the cryptosystem ROLLO-I-128 with parameters $(79, 94, 47, 5)$ needs 2^{73} bit operations to decode an error, instead of the announced 2^{128} bits of security. In the same way, ROLLO-I-192

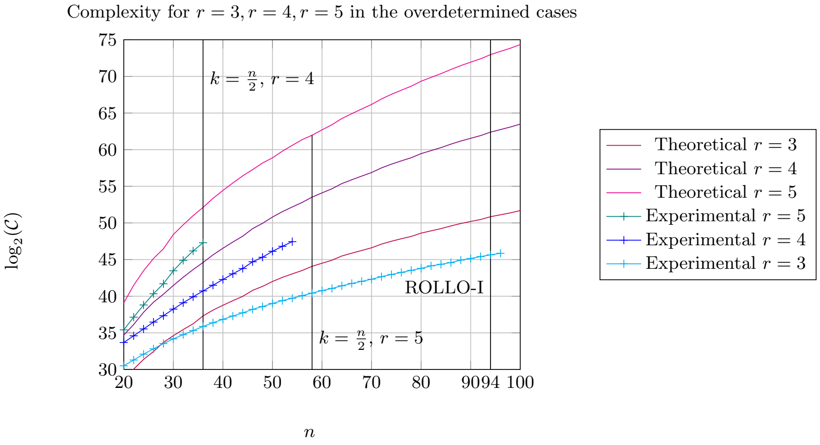


Fig. 1. Theoretical vs Experimental value of the complexity of the computation. The computations are done using `magma v2.22-2` on a machine with a Intel® Xeon® 2.00GHz processor (*Any mention of commercial products is for information only and does not imply endorsement by NIST*). We measure the experimental complexity in terms of clock cycles of the CPU, given by the `magma` function `ClockCycles()`. The theoretical value is the binary logarithm of $m\binom{n-k-1}{r}\binom{n}{r}^{2.81-1}$. The experimental values are the binary logaithms of the aforementionned experimental complexity m is the largest prime less than $n/1.18$, and k the minimum of $n/2$ (right part of the graph) and the largest value for which the system is overdetermined (left part).

with parameters $(89, 106, 53, 6)$ would have 86 bits of security instead of 192. The parameters $(113, 134, 67, 7)$ for ROLLO-I-256 are not in the overdetermined case.

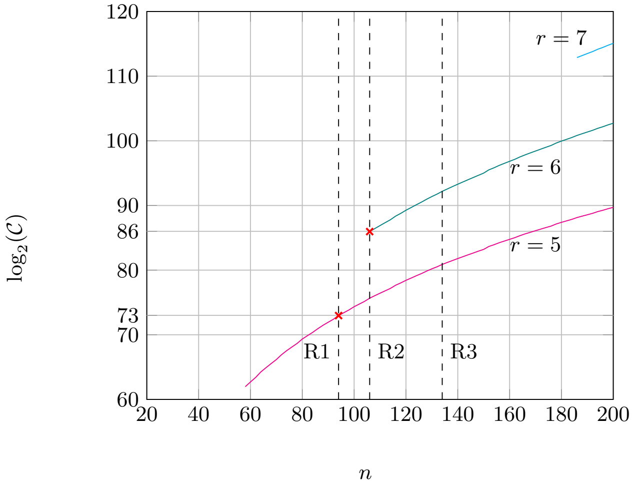
There are two classical improvements that can be used to lower the complexity of solving an algebraic system. The first one consists in selecting a subset of all equations, when some of them are redundant, see Section 4.2. The second one is the hybrid attack that will be explained in Section 4.3.

4.2 Improvement in the “super”-overdetermined case by puncturing

We consider the case when the system is “super”-overdetermined, i.e. when the number of rows in **MaxMin** is really larger than the number of columns. In that case, it is not necessary to consider all equations, we just need the minimum number of them to be able to find the solution.

To select the good equations (i.e. the ones that are likely to be linearly independent), we can take the system **MaxMinors** obtained by considering code $\widetilde{C}$ punctured on the p last coordinates, instead of the entire code. Puncturing code $\widetilde{C}$ is equivalent to shortening the dual code, i.e. considering the system

$$\textbf{MaxMinors}\left(\boldsymbol{C}_{*,\{1..n-p\}}(\boldsymbol{H}^\intercal)_{\{1..n-p\},\{1..n-k-1-p\}}\right). \tag{17}$$



Theoretical complexity for $r = 5, 6, 7$ in the *overdetermined* cases when $n = 2k$.

Fig. 2. Theoretical value of the complexity of the computation in the overdetermined cases, which is the binary logarithm of $m\binom{n-k-1}{r}\binom{n}{r}^{2.81-1}$. m is the largest prime less than $n/1.18$, $n = 2k$. The axis "R1, R2, R3" correspond to the values of n for the cryptosystems ROLLO-I-128; ROLLO-I-192 and ROLLO-I-256.

as we take $\boldsymbol{H}$ is systematic form on the last coordinates. This system is formed by a sub-sequence of polynomials in **MaxMinors** that do not contains the variables $c_{i,j}$ with $n-p+1 \leq j \leq n$. This system contains $m\binom{n-p-k-1}{r}$ equations in $\binom{n-p}{r}$ variables $\boldsymbol{C}_{*,T}$ with $T \subset \{1..n-p-k-1\}$. If we take the maximal value of p such that $m\binom{n-p-k-1}{r} \geq \binom{n-p}{r} - 1$, we can still apply Algorithm 1 but the complexity is reduced for instance to

$$\mathcal{O}\left(m\binom{n-p-k-1}{r}\binom{n-p}{r}^{\omega-1}\right) \tag{18}$$

operations in the field $\mathbb{F}_q$.

4.3 Reducing to the overdetermined case: hybrid attack

Another classical improvement consists in using an hybrid approach mixing exhaustive search and linear resolution, like in [12]. This consists in specializing some variables of the system to reduce an underdetermined case to an overdetermined one.

For instance, if we specialize a columns of the matrix $\boldsymbol{C}$, we are left with solving q^{ar} linear systems **MaxMin** of size $m\binom{n-k-1}{r} \times \binom{n-a}{r}$, and the global

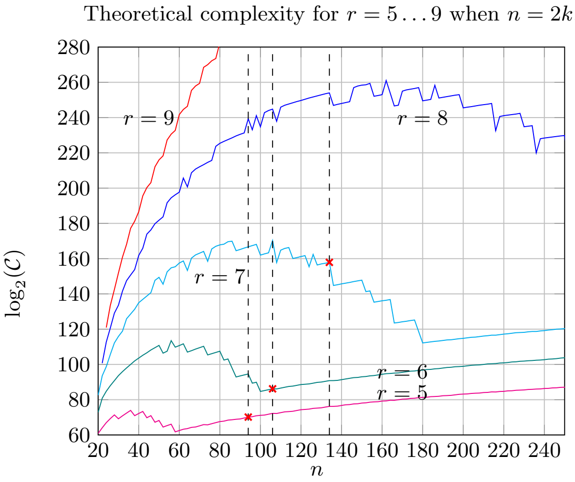


Fig. 3. Theoretical value of the complexity of RD in the overdetermined case (using punctured codes or specialization). $\mathcal{C}$ is the smallest value between (19) and (18). m is the largest prime less than $n/1.18$, $n = 2k$. The dashed axes correspond to the values of n for the cryptosystems ROLLO-I-128; ROLLO-I-192 and ROLLO-I-256.

cost is

$$\mathcal{O}\left(q^{ar} m \binom{n-k-1}{r} \binom{n-a}{r}^{\omega-1}\right) \tag{19}$$

operations in the field $\mathbb{F}_q$. In order to minimize the previous complexity (19), one chooses a to be the smallest integer such that the condition $m\binom{n-k-1}{r} \geq \binom{n-a}{r} - 1$ is fulfilled. Figure 3 page 17 gives the best theoretical complexities obtained for $r = 5 \ldots 9$ with the best values of a and p, for $n = 2k$. Table 1 page 26 gives the complexities of our attack (column "This paper") for all the parameters in the ROLLO and RQC submissions to the NIST competition; for the sake of clarity, we give the previous complexity from [11].

5 Solving Rank Decoding and MinRank problems: underdetermined case

This section analyzes the support minors modeling approach (Modeling 1).

5.1 Solving (3) by direct linearization

The number of monomials that can appear in Modeling 1 is $K\binom{n}{r}$ whereas the number of equations is $m\binom{n}{r+1}$. When the solution space of (3) is of dimension

1, we expect to solve it by direct linearization whenever:

$$m\binom{n}{r+1} \geq K\binom{n}{r} - 1. \tag{20}$$

We did a lot of experiments as explained in Section 5.6, and they suggest that it is the case.

Remark 2. Note that, in what follows, the Eq. (20) will sometimes be referred as the "$b = 1$ case".

5.2 Solving Support Minors Modeling at a higher degree

In the case where Eq. (20) does not hold we may produce a generalized version of Support Minors Modeling, multiplying the Support Minors Modeling equations by homogeneous degree $b - 1$ monomials in the linear variables, resulting in a system of equations that are homogeneous degree 1 in the variables c_T and homogeneous degree b in the variables x_i. The strategy will again be to linearize over monomials. The most common cases are $q = 2$ and $q > b$. In the former case there are $\sum_{i=1}^{b}\binom{n}{r}\binom{K}{i}$ monomials, and in the latter case there are $\binom{n}{r}\binom{K+b-1}{b}$. For the time being, we will focus on the simpler $q > b$ case. There is however an unavoidable complication which occurs whenever we consider $b \geq q$. Unlike in the simpler $b = 1$ case, for $b \geq 2$ we cannot assume that all $m\binom{n}{r+1}\binom{K+b-2}{b-1}$ equations we produce in this way are linearly independent up to the point where we can solve the system by linearization. In fact, we can construct explicit linear relations between the equations starting at $b = 2$.

This comes from determinantal identities involving maximal minors of matrices whose first rows are some of the $\boldsymbol{r_j}$'s concatenated with $\boldsymbol{C}$. For instance we may write the trivial identity for any subset J of columns of size $r + 2$:

$$\begin{vmatrix} \boldsymbol{r_j} \\ \boldsymbol{r_k} \\ \boldsymbol{C} \end{vmatrix}_{*,J} + \begin{vmatrix} \boldsymbol{r_k} \\ \boldsymbol{r_j} \\ \boldsymbol{C} \end{vmatrix}_{*,J} = 0.$$

Notice that this gives trivially a relation between certain equations corresponding to $b = 2$ since a cofactor expansion along the first row of $\begin{vmatrix} \boldsymbol{r_j} \\ \boldsymbol{r_k} \\ \boldsymbol{C} \end{vmatrix}_{*,J}$ shows that this maximal minor is indeed a linear combination of terms which is the multiplication of a linear variable x_i with a maximal minor of the matrix $\begin{pmatrix} \boldsymbol{r_k} \\ \boldsymbol{C} \end{pmatrix}$ (in other words an equation corresponding to $b = 2$). A similar result holds for $\begin{vmatrix} \boldsymbol{r_k} \\ \boldsymbol{r_j} \\ \boldsymbol{C} \end{vmatrix}_{*,J}$ where a cofactor expansion along the first row yields terms formed by a linear variable x_i multiplied by a maximal minor of the matrix $\begin{pmatrix} \boldsymbol{r_j} \\ \boldsymbol{C} \end{pmatrix}$. This result can be generalized by considering symmetric tensors $(S_{j_1,\cdots,j_r})_{\substack{1 \leq j_1 \leq m \\ \cdots \\ 1 \leq j_r \leq m}}$ of dimension m of rank $b \geq 2$ over $\mathbb{F}_q$. Recall that these are tensors that satisfy

$$S_{j_1,\cdots,j_b} = S_{j_{\sigma(1)},\cdots,j_{\sigma(b)}}$$

for any permutation σ acting on $\{1..b\}$. This is a vector space that is clearly isomorphic to the space of homogeneous polynomials of degree b in $y_1, \cdots, y_m$ over $\mathbb{F}_q$. The dimension of this space is therefore $\binom{m+b-1}{b}$. We namely have

Proposition 6. *For any symmetric tensor* $(S_{j_1,\cdots,j_b})_{\substack{1\leq j_1 \leq m\\ \cdots \\ 1 \leq j_b \leq m}}$ *of dimension* m *of rank* $b \geq 2$ *over* $\mathbb{F}_q$ *we have*

$$\sum_{j_1=1}^{m} \cdots \sum_{j_b=1}^{m} S_{j_1,\cdots,j_b} \left| \begin{matrix} \boldsymbol{r_{j_1}} \\ \cdots \\ \boldsymbol{r_{j_b}} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} = 0$$

where J *is any subset of* $\{1..n\}$ *of size* $r+b$.

Proof. Notice first that the maximal minor $\left| \begin{matrix} \boldsymbol{r_{j_1}} \\ \cdots \\ \boldsymbol{r_{j_b}} \\ \boldsymbol{C} \end{matrix} \right|_{*,J}$ is equal to 0 whenever at least two of the j_i's are equal. The left-hand sum reduces therefore to a sum of terms of the form $\sum_{\sigma \in S_b} S_{\sigma(j_1),\cdots,\sigma(j_b)} \left| \begin{matrix} \boldsymbol{r_{\sigma(j_1)}} \\ \cdots \\ \boldsymbol{r_{\sigma(j_b)}} \\ \boldsymbol{C} \end{matrix} \right|_{*,J}$ where all the j_i's are different. Notice now that from the fact that S is a symmetric tensor we have

$$\begin{aligned} \sum_{\sigma \in S_b} S_{\sigma(j_1),\cdots,\sigma(j_b)} \left| \begin{matrix} \boldsymbol{r_{\sigma(j_1)}} \\ \cdots \\ \boldsymbol{r_{\sigma(j_b)}} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} &= S_{j_1,\cdots,j_b} \sum_{\sigma \in S_b} \left| \begin{matrix} \boldsymbol{r_{\sigma(j_1)}} \\ \cdots \\ \boldsymbol{r_{\sigma(j_b)}} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} \\ &= 0 \end{aligned}$$

because the determinant is an alternating form and there as many odd and even permutations in the symmetric group of order b when $b \geq 2$. □

This proposition can be used to understand the dimension D of the space of linear equations we obtain after linearizing the equations we obtain for a certain b. For instance for $b = 2$ we obtain $m\binom{n}{r+1}K$ linear equations (they are obtained by linearizing the equations resulting from multiplying all the equations of the support minors modeling by one of the K linear variables). However as shown by Proposition 6 all of these equations are not independent and we have $\binom{n}{r+2}\binom{m+1}{2}$ linear relations coming from all relations of the kind

$$\sum_{j=1}^{m} \sum_{k=1}^{m} S_{j,k} \left| \begin{matrix} \boldsymbol{r_j} \\ \boldsymbol{r_k} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} = 0. \tag{21}$$

In our experiments, these relations turnt out to be independent yielding that the dimension D of this space should not be greater than $m\binom{n}{r+1}K - \binom{n}{r+2}\binom{m+1}{2}$. Experimentally, we observed that we indeed had

$$\mathrm{D}_{\mathrm{exp}} = m\binom{n}{r+1}K - \binom{n}{r+2}\binom{m+1}{2}.$$

For larger values of b things get more complicated but again Proposition 6 plays a key role here. Consider for example the case $b = 3$. We have in this case

$m\binom{n}{r+1}\binom{K+1}{2}$ equations obtained by multiplying all the equations of the support minors modeling by monomials of degree 2 in the linear variables. Again these equations are not all independent, there are $\binom{m+1}{2}\binom{n}{r+2}K$ equations obtained by mutiplying all the linear relations between the $b = 2$ equations derived from (21) by a linear variable, they are of the form

$$x_i \sum_{j=1}^{m}\sum_{k=1}^{m} S_{j,k} \left| \begin{matrix} \boldsymbol{r_j} \\ \boldsymbol{r_k} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} = 0. \qquad (22)$$

But all these linear relations are themselves not independent as can be checked by using Proposition 6 with $b = 3$, we namely have for any symmetric tensor S_{jkl} of rank 3:

$$\sum_{j=1}^{m}\sum_{k=1}^{m} S_{i,j,k} \left| \begin{matrix} \boldsymbol{r_i} \\ \boldsymbol{r_j} \\ \boldsymbol{r_k} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} = 0. \qquad (23)$$

This induces linear relations among the equations (22), as can be verified by a cofactor expansion along the first row of the left-hand term of (23) which yields an equation of the form

$$\sum_{i=1}^{m} x_i \sum_{j=1}^{m}\sum_{k=1}^{m} S^i_{j,k} \left| \begin{matrix} \boldsymbol{r_j} \\ \boldsymbol{r_k} \\ \boldsymbol{C} \end{matrix} \right|_{*,J} = 0$$

where the $\boldsymbol{S}^i = (S^i_{j,k})_{\substack{1\leq j\leq m \\ 1\leq k\leq m}}$ are symmetric tensors of order 2. We would then expect that the dimension of the set of linear equations obtained from (22) is only $\binom{m+1}{2}\binom{n}{r+2}K - \binom{n}{r+3}\binom{m+2}{3}$ yielding an overall dimension D of the linearized system of order

$$\mathrm{D} = m\binom{n}{r+1}\binom{K+1}{2} - \binom{m+1}{2}\binom{n}{r+2}K + \binom{n}{r+3}\binom{m+2}{3},$$

which is precisely what we observe experimentally. This argument extends also to higher values of b, so that, if linear relations of the form considered above are the only relevant linear relations, then the number of linearly independent equations available for linearization at a given value of b is:

Heuristic 2

$$\mathrm{D}_{\mathrm{exp}} = \sum_{i=1}^{b} (-1)^{i+1} \binom{n}{r+i}\binom{m+i-1}{i}\binom{K+b-i-1}{b-i}. \qquad (24)$$

Experimentally, we found this to be the case with overwhelming probability (see Section 5.6) with the only general exceptions being:

1. When $\mathrm{D}_{\mathrm{exp}}$ exceeds the number of monomials for a smaller value of b, typically 1, the number of equations is observed to be equal to the number of monomials for all higher values of b as well, even if $\mathrm{D}_{\mathrm{exp}}$ does not exceed the total number of monomials at these higher values of b.

2. When the underlying MinRank Problem has a nontrivial solution and cannot be solved a $b = 1$, we find the maximum number of linearly independent equations is not the total number of monomials but is less by 1. This is expected, since when the underlying MinRank problem has a nontrivial solution, then the Support Minors Modeling equations have a 1 dimensional solution space.
3. When $b \geq r + 2$, the equations are not any more linearly independent, and we give an explanation in Section 5.4.

In summary, in the general case, we expect to be able to linearize at degree b whenever $b < r + 2$ and

$$\binom{n}{r}\binom{K+b-1}{b} - 1 \leq \sum_{i=1}^{b}(-1)^{i+1}\binom{n}{r+i}\binom{m+i-1}{i}\binom{K+b-i-1}{b-i} \quad (25)$$

Note that, for $b = 1$, we recover the result (20). As this system is very sparse, with $K(r + 1)$ monomials per equation, one can solve it using Wiedemann algorithm [39]; thus the complexity to solve MinRank problem is

$$\mathcal{O}\left(K(r+1)\left(\binom{n}{r}\binom{K+b-1}{b}\right)^2\right) \quad (26)$$

where b is the smallest positive integer so that the condition (25) is fulfilled.

5.3 The $q = 2$ case

The same considerations apply in the $q = 2$ case, but due to the field equations, $x_i^2 = x_i$, for systems with $b \geq 2$, a number of monomials will collapse to a lower degree. This results in a system which is no longer homogeneous. Thus, in this case it is most profitable to combine the equations obtained at a given value of b with those produced using all smaller values of b. Similar considerations to the general case imply that as long as $b < r + 2$ we will have

$$\mathrm{D}_{\mathrm{exp}} = \sum_{j=1}^{b}\sum_{i=1}^{j}(-1)^{i+1}\binom{n}{r+i}\binom{m+i-1}{i}\binom{K}{j-i}. \quad (27)$$

equations with which to linearize the

$$\sum_{j=1}^{b}\binom{n}{r}\binom{K}{j}$$

monomials that occur at a given value of b. We therefore expect to be able to solve by linearization when $b < r + 2$ and b is large enough that

$$\sum_{j=1}^{b}\binom{n}{r}\binom{K}{j} - 1 \leq \sum_{j=1}^{b}\sum_{i=1}^{j}(-1)^{i+1}\binom{n}{r+i}\binom{m+i-1}{i}\binom{K}{j-i}. \quad (28)$$

Similarly to the general case for any q described in the previous section, the complexity to solve MinRank problem when $q = 2$ is

$$\mathcal{O}\left(K(r+1)\left(\sum_{j=1}^{b}\binom{n}{r}\binom{K}{j}\right)^2\right) \tag{29}$$

where b is the smallest positive integer so that the condition (28) is fulfilled.

5.4 Toward the $b \geq r + 2$ case

We can also construct additional nontrivial linear relations starting at $b = r+2$. The simplest example of this sort of linear relation occurs when $m > r + 1$. Note that each of the Support Minors modeling equations at $b = 1$ is bilinear in the x_i variables and a subset consisting of $r + 1$ of the variables c_T. Note also, that there are a total of m equations derived from the same subset (one for each row of $\sum_{i=0}^{K} x_i M_i$.) Therefore, if we consider the Jacobian of the $b = 1$ equations with respect to the variables c_T, the m equations involving only $r + 1$ of the variables c_T will form a submatrix with m rows and only $r + 1$ nonzero columns. Using a Cramer-like formula, we can therefore construct left kernel vectors for these equations; its coefficients would be degree $r + 1$ polynomials in the x_i variables. Multiplying the equations by this kernel vector will produce zero, because the $b = 1$ equations are homogeneous, and multiplying equations from a bilinear system by a kernel vector of the Jacobian of that system cancels all the highest degree terms. This suggests that Eq. (24) needs to be modified when we consider values of b that are $r + 2$ or greater. These additional linear relations do not appear to be relevant in the most interesting range of b for attacks on any of the cryptosystems considered, however.

5.5 Improvements for Generic Minrank

The two classical improvements Section 4.2 in the “super”-overdetermined cases the hybrid attack and Section 4.3 can also apply for Generic Minrank.

We can consider applying the Support Minors Modeling techniques to submatrices $\sum_{i=1}^{K} \boldsymbol{M}'_i x_i$ of $\sum_{i=1}^{K} \boldsymbol{M}_i x_i$. Note that if $\sum_{i=1}^{K} \boldsymbol{M}_i x_i$ has rank less than or equal to r, so does $\sum_{i=1}^{K} \boldsymbol{M}'_i x_i$, so assuming we have a unique solution x_i to both systems of equations, it will be the same. Generically, we will keep a unique solution in the smaller system as long as the decoding problem has a unique solution, i.e. as long as the Johnson bound $K \leq (m-r)(n-r)$ is satisfied.

We generally find that the most beneficial settings use matrices with all m rows, but only $n' \leq n$ of the columns. This corresponds to a puncturing of the corresponding $\mathbb{F}_q$ matrix code. It is always beneficial for the attacker to reduce n' to the minimum value allowing linearization at a given degree b, however, it can sometimes lead to an even lower complexity to reduce n' further and solve at a higher degree b.

On the other side, we can run exhaustive search on a variables x_i in $\mathbb{F}_q$ and solve q^a systems with a smaller value of b, so that the resulting complexity is smaller than solving directly the system with a higher value of b. This optimization is considered in the attack against ROLLO-I-256 (see Table 1); more detail about this example are given in Section 6.1.

5.6 Experimental results for Generic Minrank

We verified experimentally that the value of $\mathrm{D}_{\mathrm{exp}}$ correctly predicts the number of linearly independent polynomials. We constructed random systems (with and without a solution) for $q = 2, 13$, with $m = 7, 8$, $r = 2, 3$, $n = r + 3, r + 4, r + 5$, $K = 3, \ldots, 20$. In most of the cases, the number of linearly independent polynomials was as expected. For $q = 13$, we had a few number of non-generic systems (usually less than 1% over 1000 random sampling), and only in square cases where the matrices have a predicted rank equal to the number of columns. For $q = 2$ we had a higher probability of linear dependences, due to the fact that over a such small field, random matrices have a non-trivial probability to be non invertible. Anyway, as soon as the field is big enough or the number $\mathrm{D}_{\mathrm{exp}}$ is large compared to the number of columns, 100% of our experiments succeeded over 1000 samples.

5.7 Using Support Minors Modeling in conjunction with MaxMin for RD

Recall that from MaxMin, we obtain $m\binom{n-k-1}{r}$ homogeneous linear equations in the variables c_T. These can be used to produce equations over the same monomials as used for Support Minors Modeling with $K = mk + 1$. In the $q > b$ case, this can be done by multiplying the equations from MaxMin by homogeneous degree b monomials in the variables x_i. In the $q = 2$ case this can be done by multiplying the MaxMin equations by monomials of degree b or less. With all the arguments mentioned above and the experiments mentioned in Section 5.6, we can make a similar heuristic as Heuristic 1, this suggests that linearization is possible for $q > b$, $0 < b < r + 2$ whenever:

$$\binom{n}{r}\binom{mk+b}{b} - 1 \leq$$
$$m\binom{n-k-1}{r}\binom{mk+b}{b} + \sum_{i=1}^{b}(-1)^{i+1}\binom{n}{r+i}\binom{m+i-1}{i}\binom{mk+b-i}{b-i}, \tag{30}$$

and for $q = 2$, $0 < b < r + 2$ whenever:

$$A_b - 1 \leq B_c + C_b \tag{31}$$

where

$$A_b := \sum_{j=1}^{b} \binom{n}{r} \binom{mk+1}{j}$$

$$B_b := \sum_{j=1}^{b} \left(m \binom{n-k-1}{r} \binom{mk+1}{j} \right)$$

$$C_b := \sum_{j=1}^{b} \sum_{i=1}^{j} \left((-1)^{i+1} \binom{n}{r+i} \binom{m+i-1}{i} \binom{mk+1}{j-i} \right).$$

For the latter, it leads to a complexity of

$$\mathcal{O}\left((B_b + C_b) A_b^{\omega-1}\right) \tag{32}$$

where b is the smallest positive integer so that the condition (31) is fulfilled. This complexity formula correspond to solving a linear system with A_b unknowns and $B_b + C_b$ equations, recall that ω is the constant of linear algebra.

One notices that for a large range of parameters, this system is particularly sparse, so one could take advantage of that to use Wiedemann algorithm [39]. More precisely, for values of m, n, r and k of ROLLO or RQC parameters (see Table 4 and Table 5) for which the condition (31) is fulfilled, we typically find that $b \approx r$.

In this case, B_b equations consist of $\binom{k+r+1}{r}$ monomials, C_b equations consist of $(mk+1)(r+1)$ monomials, and the total space of monomials is of size A_b. The Wiedemann's algorithm complexity can be written in term of the average number of monomials per equation, in our case it is

$$\frac{B_b \binom{k+r+1}{r} + C_b (mk+1)(r+1)}{B_b + C_b}.$$

Thus the linearized system at degree b is sufficiently sparse that Wiedemann outperforms Strassen for $b \geq 2$. Therefore the complexity of support minors modeling bootstrapping MaxMin for RD is

$$\mathcal{O}\left(\frac{B_b \binom{k+r+1}{r} + C_b (mk+1)(r+1)}{B_b + C_b} A_b^2 \right) \tag{33}$$

where b is still the smallest positive integer so that the condition (31) is fulfilled. A similar formula applies for the case $q > b$.

5.8 Last step of the attack

To end the attack on MinRank using Support Minors modeling or the attack on Rank Decoding using MaxMinors modeling in conjunction with Support Minors modeling, one needs to find the affectations for each unknowns. Indeed, unlike the case where one uses only MaxMinors modeling with the specialization in $\boldsymbol{C}$ (see Section 4), the direct linearization does not lead to affectations of the form $c_{i,j} = \mu_{i,j}$ where $\mu_{i,j}$ are elements in $\mathbb{F}_q$. In fact, with the Support Minors modeling one gets affectations of the form

$$x^\alpha c_{i,j} = \mu \in \mathbb{F}_q$$

where the x^α's are monomials of degree $b-1$ in the x_i's variables.

In order to extract the values of all the x_i's and thus finish the attack, one needs to specialize $x_1 = 1$, this is possible as long as $x_1 \neq 0$ since the solution space has dimension 1; if this specialization does not lead to a unique solution, one tries with x_2 and so on. Then, one computes quotients of the form

$$\frac{x_1}{x_l} = \frac{x_1 x^\alpha c_{i_0,j_0}}{x_l x^\alpha c_{i_0,j_0}}, \quad \deg(x^\alpha) = b-2 \tag{34}$$

for all the values of l in $\{2..K\}$ with a fixed minor c_{i_0,j_0}. Doing so, one gets the values of all the x_i's and thus finish the attack. This works as long as the minor c_{i_0,j_0} of $\boldsymbol{C}$ chosen is non-zero, if it is, one uses another one, and so on; our experiments always worked with one or two minors.

6 Complexity of the attacks for different cryptosystems and comparison with generic Gröbner basis approaches

6.1 Attacks against the Rank Decoding problem

Table 1 presents the best complexity of our attacks (see sections 4 and 5) against RD and gives the binary logarithm of the complexities (column "This paper") for all the parameters in the ROLLO and RQC submissions to the NIST competition and Loidreau cryptosystem [30]; for the sake of clarity, we give the previous best known complexity from [11] (last column). The third column gives the original rate for being overdeterminate. The column 'a' indicates the number of specialized columns in the hybrid approach (Section 4.3), when the system is not overdetermined. Column 'p' indicates the number of punctured columns in the "super"-overdetermined cases (Section 4.2). Column 'b' indicates that we use Support Minors Modeling in conjunction with MaxMin (Section 5.7).

Let us give more detail on the way to compute the best complexity for ROLLO-I-256 in Table 1. Recall that its parameters are $(m, n, k, r) = (113, 134, 67, 7)$. The attack from Section 4 only works with the hybrid approach, thus requiring $a = 8$ and resulting in a complexity of 158 bits (using (19) and $\omega = 2.81$). On the other hand, the attack from Section 5.7 needs $b = 2$ which results in a complexity of 154 (this time using Wiedemann's algorithm). However, if we specialize $a = 3$ columns in $\boldsymbol{C}$, we get $b = 1$ and the resulting complexity using Wiedemann's algorithm is 151.

Table 1. Complexity of the attack against Rank Decoding for different cryptosystems. The values in the column **This paper** are the smallest ones between Strassen's and Wiedemann's algorithm, the "*" indicates that it is Wiedemann.

	(m,n,k,r)	$\frac{m\binom{n-k-1}{r}}{\binom{n}{r}-1}$	a	p	b	**This paper**	[11]
Loidreau ([30])	$(128,120,80,4)$	1.28	0	43	0	**65**	98
ROLLO-I-128	$(79,94,47,5)$	1.97	0	9	0	**71**	117
ROLLO-I-192	$(89,106,53,6)$	1.06	0	0	0	**87**	144
ROLLO-I-256	$(113,134,67,7)$	0.67	3	0	1	**151***	197
ROLLO-II-128	$(83,298,149,5)$	2.42	0	40	0	**93**	134
ROLLO-II-192	$(107,302,151,6)$	1.53	0	18	0	**111**	164
ROLLO-II-256	$(127,314,157,7)$	0.89	0	6	1	**159***	217
ROLLO-III-128	$(101,94,47,5)$	2.52	0	12	0	**70**	119
ROLLO-III-192	$(107,118,59,6)$	1.31	0	4	0	**88**	148
ROLLO-III-256	$(131,134,67,7)$	0.78	0	0	1	**131***	200
RQC-I	$(97,134,67,5)$	2.60	0	18	0	**77**	123
RQC-II	$(107,202,101,6)$	1.46	0	10	0	**101**	156
RQC-III	$(137,262,131,7)$	0.93	3	0	0	**144**	214

6.2 Attacks against the MinRank problem

Tables 2 and 3 show the complexity of our attack against generic MinRank problem for GeMSS and Rainbow, two cryptosystems at the second round of the aforementioned NIST competition. The two tables also compare this new attack to the previous MinRank attacks, which use minors modeling in the case of GeMSS [14] and a linear algebra search [18] in the case of Rainbow. In table 3, the column "Best/Type" shows the complexity of the current best attack against Rainbow, which is not a MinRank attack.

6.3 Comparison between our approach and the use of generic Gröbner basis algorithms

Since our approach is an algebraic attack, it relies on solving a polynomial system, thus it does look like a Gröbner basis computation. In fact, we do compute a Gröbner basis of the system, as we compute the unique solution of the system, which represents its Gröbner basis.

Nevertheless, our algorithm is not a generic Gröbner basis algorithm as it only works for the special type of system studied in this paper: the RD and MinRank systems. As it is specifically designed for this purpose and for the reasons detailed below, it is more efficient than a generic algorithm.

There are three main reasons why our approach is more efficient than a generic Gröbner basis algorithm:

- We compute formally (that is to say at no extra cost except the size of the equations) new equations of degree r (the **MaxMinors** ones) that are

Table 2. Complexity comparison between the new and the previous MinRank attacks against GeMSS parameters. Recall that the previous attack used minors (see [14]). The new complexity is computed by finding the number of columns n' and the degree b that minimizes the complexity, as described in Section 5.

						Complexity	
(D, n, Δ, v)	n/m	K	r	n'	b	New	Previous
GeMSS128$(513, 174, 12, 12)$	174	162	34	61	2	**154**	522
GeMSS192$(513, 256, 22, 20)$	265	243	52	94	2	**223**	537
GeMSS256$(513, 354, 30, 33)$	354	324	73	126	3	**299**	1254
RedGeMSS128$(17, 177, 15, 15)$	177	162	35	62	2	**156**	538
RedGeMSS192$(17, 266, 23, 25)$	266	243	53	95	2	**224**	870
RedGeMSS256$(17, 358, 34, 35)$	358	324	74	127	3	**301**	1273
BlueGeMSS128$(129, 175, 13, 14)$	175	162	35	63	2	**158**	537
BlueGeMSS192$(129, 265, 22, 23)$	265	243	53	95	2	**224**	870
BlueGeMSS256$(129, 358, 34, 32)$	358	324	74	127	3	**301**	1273

Table 3. Comparison between the new MinRank attack, the previous best MinRank attack using linear algebra search, and the best known attack for Rainbow. Here the acronyms RBS and DA stand from Rainbow Band Separation and Direct Algebraic, respectively [18]. The new complexity is computed by finding the number of columns n' and the degree b that minimizes the complexity, as described in Section 5.

						Complexity		
Rainbow$(GF(q), v_1, o_1, o_2)$	n	K	r	n'	b	New	Previous	Best / Type
Ia$(GF(16), 32, 32, 32)$	96	33	64	82	3	155	161	145/RBS
IIIc$(GF(256), 68, 36, 36)$	140	37	104	125	5	**208**	585	215/DA
Vc$(GF(256), 92, 48, 48)$	188	49	140	169	5	**272**	778	275/DA

already in the ideal, but not in the vector space

$\mathcal{F}_r := \langle uf : u$ monomial of degree $r-2$, f in the set of initial polynomials$\rangle$.

In fact, a careful analysis of a Gröbner basis computation with a normal strategy shows that those equations are in $\mathcal{F}_{r+1}$, and that the first degree fall for those systems is $r+1$. Here, we apply linear algebra directly on a small number of polynomials of degree r (see the two following items for more details), whereas a generic Gröbner basis algorithm would compute a lot of polynomials of degree $r+1$ and then reduce them in order to get those polynomials of degree r.

- A classical Gröbner basis algorithm using linear algebra and a normal strategy will construct matrices like the Macaulay ones, where the rows correspond to polynomials in the ideal and the columns to monomials of a certain degree. Here, we introduce variables c_T that represent maximal minors of $\boldsymbol{C}$, and thus represent not one monomial of degree r, but $r!$ monomials of degree r. As we compute the Gröbner basis by using only polynomials that can be expressed in terms of those variables (see the last item below), this reduces the number of columns of our matrices by a factor around $r!$ compared to generic Macaulay-like matrices.
- The solution can be found by applying linear algebra only to some specific equations, namely the MaxMinors ones in the overdetermined case, and in the underdetermined case, equations that have degree 1 in the c_T variables, and degree $b-1$ in the x_i variables (see Section 5.2). This enables us to deal with polynomials involving only the c_T variables and the x_i variables, whereas a generic Gröbner basis algorithm would consider all monomials up to degree $r+b$ in the x_l and the $c_{i,j}$ variables. This drastically reduces the number of rows and columns in our matrices.

For all of those reasons, in the overdetermined case, only an elimination on our selected MaxMinors equations (with a "compacted" matrix with respect to the columns) is sufficient to get the solution; so we essentially avoid going up to the degree $r+1$ to produce those equations, we select a small number of rows, and gain a factor $r!$ on the number of columns.

In the underdetermined case, we find linear equations by linearization on some well-chosen subspaces of the vector space $\mathcal{F}_{r+b}$. We have theoretical reasons to believe that our choice of subspace should lead to the computation of the solution (as usual, this is a "genericity" hypothesis), and it is confirmed by all our experiments.

7 Examples of new parameters for ROLLO-I and RQC

In light of the attacks presented in this article, it is possible to give a few examples of new parameters for the rank-based cryptosystems, submitted to the NIST competition, ROLLO and RQC. With these new parameters, ROLLO and RQC would be resistant to our attacks, while still remaining attractive, for example with a loss of only about 50 % in terms of key size for ROLLO-I.

For cryptographic purpose, parameters have to belong to an area which does not correspond to the overdetermined case and such that the hybrid approach would make the attack worse than in the underdetermined case.

Alongside the algebraic attacks in this paper, the best combinatorial attack against RD is in [4]; as a reminder, for attacking a $[n, k]$ code over $\mathbb{F}_{q^m}$ with target rank r, its complexity is

$$\mathcal{O}\left((nm)^2 q^{r\left\lceil \frac{m(k+1)}{n} \right\rceil - m}\right).$$

Remark 3. In this section, the notation is chosen to match the one in ROLLO and RQC submissions' specifications ([7] and [1]). One should be careful that here, n is the block-length and not the length of the code which can be either $2n$ or $3n$.

In what follows, we consider $\omega = 2.81$ and we use the following notation, for ROLLO (Table 4):

- **over/hybrid** is the cost of the hybrid attack; the value of a is the smallest to reach the overdetermined case, $a = 0$ means that parameters are already in the overdetermined case,
- **under** is the case of underdetermined attack.
- **comb** is the the cost of the best combinatorial attack mentioned above,
- **DFR** is the binary logarithm of the Decoding Failure Rate,

and for RQC (Table 5):

- **hyb2n(a)**: hybrid attack for length $2n$, the value of a is the smallest to reach the overdetermined case, $a = 0$ means that parameters are already in the overdetermined case,
- **hyb3n(a)**: non-homogeneous hybrid attack for length $3n$, a is the same as above. This attack corresponds to an adaptation of our attack to a non-homogeneous error of the RQC scheme, more details are given in [1],
- **und2n**: underdetermined attack for length $2n$,
- **comb3n**: combinatorial attack for length $3n$.

For more details about those parameters and the aforementioned attacks, reader may refer to the submissions specifications of ROLLO (see [7]) and RQC (see [1]).

Instance	q	n	m	r	d	pk size (B)	DFR	over/hybrid	a	p	under	b	comb
new2ROLLO-I-128	2	83	73	7	8	757	-27	233	18	0	180	3	213
new2ROLLO-I-192	2	97	89	8	8	1057	-33	258*	17	0	197*	3	283*
new2ROLLO-I-256	2	113	103	9	9	1454	-33	408*	30	0	283*	6	376*

Table 4. New parameters and attacks complexities for ROLLO-I.

Instance	q	n	m	k	w	w_r	δ	pk (B)	hyb2n(a)	hyb3n(a)	und2n	b	comb3n
newRQC-I	2	113	127	3	7	7	6	1793	160(6)	211(0)	158	1	205
newRQC-II	2	149	151	5	8	8	8	2812	331(24)	262(0)	224	3	289
newRQC-III	2	179	181	3	9	9	7	4049	553(44)	321(5)	324	6	401

Table 5. New parameters and attacks complexities for RQC.

8 Conclusion

In this paper we improve on the results by [11] on the Rank Decoding problem by providing a better analysis which permits to avoid the use of generic Gröbner basis algorithms and permits to completely break rank-based cryptosystems parameters proposed to the NIST Standardization Process, when analysis in [11] only attacked slightly these parameters (mostly corresponding to the overdeterminate case defined in [11]).

We generalize this approach to the case of the MinRank problem for which we obtain the best known complexity with algebraic attacks. We also proposed a new approach for the underdeterminate case as described in [11], for some parameters this attack supersedes the results of [11], in particular for attacking ROLLO-I-256 parameters.

Overall the results proposed in this paper give a new and deeper understanding of the complexity of difficult problems based on the rank metric. These problems have a strong interest since many systems still in the second round of the NIST standardization process, like ROLLO, RQC, GeMSS or Rainbow can be attacked through these problems.

Acknowledgements

This work has been supported by the French ANR projects CBCRYPT (ANR-17-CE39-0007) and the MOUSTIC project with the support from the European Regional Development Fund (ERDF) and the Regional Council of Normandie.

Javier Verbel was supported for this work by Colciencias scholarship 757 for PhD studies and the University of Louisville facilities.

We would like to thank John B Baena, and Karan Khathuria for useful discussions. We thank the Facultad de Ciencias of the Universidad Nacional de Colombia sede Medellín for granting us access to the Enlace server, where we ran some of the experiments.

References

1. Aguilar Melchor, C., Aragon, N., Bettaieb, S., Bidoux, L., Blazy, O., Bros, M., Couvreur, A., Deneuville, J.C., Gaborit, P., Hauteville, A., Zémor, G.: Rank quasi cyclic (RQC). Second round submission to the NIST post-quantum cryptography call (Apr 2020), https://pqc-rqc.org

2. Aguilar Melchor, C., Aragon, N., Bettaieb, S., Bidoux, L., Blazy, O., Deneuville, J.C., Gaborit, P., Hauteville, A., Zémor, G.: Ouroboros-R. First round submission to the NIST post-quantum cryptography call (Nov 2017), https://pqc-ouroborosr.org
3. Aguilar Melchor, C., Aragon, N., Bettaieb, S., Bidoux, L., Blazy, O., Deneuville, J.C., Gaborit, P., Zémor, G.: Rank quasi cyclic (RQC). First round submission to the NIST post-quantum cryptography call (Nov 2017), https://pqc-rqc.org
4. Aragon, N., Gaborit, P., Hauteville, A., Tillich, J.P.: A new algorithm for solving the rank syndrome decoding problem. In: Proc. IEEE ISIT (2018)
5. Aragon, N., Blazy, O., Deneuville, J.C., Gaborit, P., Hauteville, A., Ruatta, O., Tillich, J.P., Zémor, G.: LAKE – Low rAnk parity check codes Key Exchange. First round submission to the NIST post-quantum cryptography call (Nov 2017), https://csrc.nist.gov/CSRC/media/Projects/Post-Quantum-Cryptography/documents/round-1/submissions/LAKE.zip
6. Aragon, N., Blazy, O., Deneuville, J.C., Gaborit, P., Hauteville, A., Ruatta, O., Tillich, J.P., Zémor, G.: LOCKER – LOw rank parity CheckK codes EncRyption. First round submission to the NIST post-quantum cryptography call (Nov 2017), https://csrc.nist.gov/CSRC/media/Projects/Post-Quantum-Cryptography/documents/round-1/submissions/LOCKER.zip
7. Aragon, N., Blazy, O., Deneuville, J.C., Gaborit, P., Hauteville, A., Ruatta, O., Tillich, J.P., Zémor, G., Aguilar Melchor, C., Bettaieb, S., Bidoux, L., Bardet, M., Otmani, A.: ROLLO (merger of Rank-Ouroboros, LAKE and LOCKER). Second round submission to the NIST post-quantum cryptography call (Apr 2020), https://pqc-rollo.org
8. Aragon, N., Blazy, O., Gaborit, P., Hauteville, A., Zémor, G.: Durandal: a rank metric based signature scheme. In: Advances in Cryptology - EUROCRYPT 2019 - 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19-23, 2019, Proceedings, Part III. LNCS, vol. 11478, pp. 728–758. Springer (2019). https://doi.org/10.1007/978-3-030-17659-4_25, https://doi.org/10.1007/978-3-030-17659-4_25
9. Aragon, N., Gaborit, P., Hauteville, A., Ruatta, O., Zémor, G.: Ranksign – a signature proposal for the NIST's call. First round submission to the NIST post-quantum cryptography call (Nov 2017), https://csrc.nist.gov/CSRC/media/Projects/Post-Quantum-Cryptography/documents/round-1/submissions/RankSign.zip
10. Aragon, N., Gaborit, P., Hauteville, A., Tillich, J.P.: A new algorithm for solving the rank syndrome decoding problem. In: 2018 IEEE International Symposium on Information Theory, ISIT 2018, Vail, CO, USA, June 17-22, 2018. pp. 2421–2425. IEEE (2018). https://doi.org/10.1109/ISIT.2018.8437464
11. Bardet, M., Briaud, P., Bros, M., Gaborit, P., Neiger, V., Ruatta, O., Tillich, J.P.: An Algebraic Attack on Rank Metric Code-Based Cryptosystems. Advances in Cryptology - EUROCRYPT 2020 (May 2020), https://arxiv.org/abs/1910.00810
12. Bettale, L., Faugere, J.C., Perret, L.: Hybrid approach for solving multivariate systems over finite fields. Journal of Mathematical Cryptology **3**(3), 177–197 (2009)
13. Buss, J.F., Frandsen, G.S., Shallit, J.O.: The computational complexity of some problems of linear algebra. J. Comput. System Sci. **58**(3), 572–596 (Jun 1999)
14. Casanova, A., Faugère, J., Macario-Rat, G., Patarin, J., Perret, L., Ryckeghem, J.: GeMSS: A Great Multivariate Short Signature. Second round submission to the NIST post-quantum cryptography call (Apr 2019), https://csrc.nist.gov/Projects/post-quantum-cryptography/round-2-submissions/GeMSS-Round2.zip

15. Courtois, N.: Efficient zero-knowledge authentication based on a linear algebra problem MinRank. In: Advances in Cryptology - ASIACRYPT 2001. LNCS, vol. 2248, pp. 402–421. Springer, Gold Coast, Australia (2001)
16. Debris-Alazard, T., Tillich, J.P.: Two attacks on rank metric code-based schemes: Ranksign and an identity-based-encryption scheme. In: Advances in Cryptology - ASIACRYPT 2018. pp. 62–92. LNCS, Springer, Brisbane, Australia (Dec 2018)
17. Ding, J., Chen, M.S., Petzoldt, A., Schmidt, D., Yang, B.Y.: Gui. First round submission to the NIST post-quantum cryptography call (Nov 2017), https://csrc.nist.gov/Projects/post-quantum-cryptography/round-2-submissions/Rainbow-Round2.zip
18. Ding, J., Chen, M.S., Petzoldt, A., Schmidt, D., Yang, B.Y.: Rainbow. Second round submission to the NIST post-quantum cryptography call (Apr 2019), https://csrc.nist.gov/Projects/post-quantum-cryptography/round-2-submissions/Rainbow-Round2.zip
19. Faugère, J.C., Levy-dit-Vehel, F., Perret, L.: Cryptanalysis of Minrank. In: Wagner, D. (ed.) Advances in Cryptology - CRYPTO 2008. LNCS, vol. 5157, pp. 280–296 (2008)
20. Faugère, J., Safey El Din, M., Spaenlehauer, P.: Computing loci of rank defects of linear matrices using Gröbner bases and applications to cryptology. In: International Symposium on Symbolic and Algebraic Computation, ISSAC 2010, Munich, Germany, July 25-28, 2010. pp. 257–264 (2010). https://doi.org/10.1145/1837934.1837984
21. Gabidulin, E.M.: Theory of codes with maximum rank distance. Problemy Peredachi Informatsii **21**(1), 3–16 (1985)
22. Gabidulin, E.M., Paramonov, A.V., Tretjakov, O.V.: Ideals over a non-commutative ring and their applications to cryptography. In: Advances in Cryptology - EUROCRYPT'91. pp. 482–489. No. 547 in LNCS, Brighton (Apr 1991)
23. Gaborit, P., Murat, G., Ruatta, O., Zémor, G.: Low rank parity check codes and their application to cryptography. In: Proceedings of the Workshop on Coding and Cryptography WCC'2013. Bergen, Norway (2013), www.selmer.uib.no/WCC2013/pdfs/Gaborit.pdf
24. Gaborit, P., Ruatta, O., Schrek, J.: On the complexity of the rank syndrome decoding problem. IEEE Trans. Information Theory **62**(2), 1006–1019 (2016)
25. Gaborit, P., Ruatta, O., Schrek, J., Zémor, G.: New results for rank-based cryptography. In: Progress in Cryptology - AFRICACRYPT 2014. LNCS, vol. 8469, pp. 1–12 (2014)
26. Gaborit, P., Ruatta, O., Schrek, J., Zémor, G.: Ranksign: An efficient signature algorithm based on the rank metric (extended version on arxiv). In: Post-Quantum Cryptography 2014. LNCS, vol. 8772, pp. 88–107. Springer (2014), https://arxiv.org/pdf/1606.00629.pdf
27. Gaborit, P., Zémor, G.: On the hardness of the decoding and the minimum distance problems for rank codes. IEEE Trans. Information Theory **62(12)**, 7245–7252 (2016)
28. Hoffstein, J., Pipher, J., Silverman, J.H.: NTRU: A ring-based public key cryptosystem. In: Buhler, J. (ed.) Algorithmic Number Theory, Third International Symposium, ANTS-III, Portland, Oregon, USA, June 21-25, 1998, Proceedings. LNCS, vol. 1423, pp. 267–288. Springer (1998)
29. Kipnis, A., Shamir, A.: Cryptanalysis of the HFE public key cryptosystem by relinearization. In: Advances in Cryptology - CRYPTO'99. LNCS, vol. 1666, pp. 19–30. Springer, Santa Barbara, California, USA (Aug 1999). https://doi.org/10.1007/3-540-48405-1

30. Loidreau, P.: A new rank metric codes based encryption scheme. In: Post-Quantum Cryptography 2017. LNCS, vol. 10346, pp. 3–17. Springer (2017)
31. Misoczki, R., Tillich, J.P., Sendrier, N., Barreto, P.S.L.M.: MDPC-McEliece: New McEliece variants from moderate density parity-check codes (2012), http://eprint.iacr.org/2012/409
32. Otmani, A., Talé-Kalachi, H., Ndjeya, S.: Improved cryptanalysis of rank metric schemes based on Gabidulin codes. Des. Codes Cryptogr. **86**(9), 1983–1996 (2018). https://doi.org/10.1007/s10623-017-0434-5, https://doi.org/10.1007/s10623-017-0434-5
33. Ourivski, A.V., Johansson, T.: New technique for decoding codes in the rank metric and its cryptography applications. Problems of Information Transmission **38**(3), 237–246 (2002). https://doi.org/10.1023/A:1020369320078
34. Overbeck, R.: A new structural attack for GPT and variants. In: Mycrypt. LNCS, vol. 3715, pp. 50–63 (2005)
35. Patarin, J.: Hidden fields equations (HFE) and isomorphisms of polynomials (IP): two new families of asymmetric algorithms. In: Maurer, U.M. (ed.) Advances in Cryptology - EUROCRYPT '96, International Conference on the Theory and Application of Cryptographic Techniques, Saragossa, Spain, May 12-16, 1996, Proceeding. LNCS, vol. 1070, pp. 33–48. Springer (1996). https://doi.org/10.1007/3-540-68339-9_4, https://doi.org/10.1007/3-540-68339-9_4
36. Petzoldt, A., Chen, M., Yang, B., Tao, C., Ding, J.: Design principles for HFEv-based multivariate signature schemes. In: Iwata, T., Cheon, J.H. (eds.) Advances in Cryptology - ASIACRYPT 2015 - 21st International Conference on the Theory and Application of Cryptology and Information Security, Auckland, New Zealand, November 29 - December 3, 2015, Proceedings, Part I. LNCS, vol. 9452, pp. 311–334. Springer (2015). https://doi.org/10.1007/978-3-662-48797-6_14, https://doi.org/10.1007/978-3-662-48797-6_14
37. Porras, J., Baena, J., Ding, J.: ZHFE, a new multivariate public key encryption scheme. In: Mosca, M. (ed.) Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. LNCS, vol. 8772, pp. 229–245. Springer (2014). https://doi.org/10.1007/978-3-319-11659-4_14, https://doi.org/10.1007/978-3-319-11659-4_14
38. Verbel, J., Baena, J., Cabarcas, D., Perlner, R., Smith-Tone, D.: On the complexity of "superdetermined" Minrank instances. In: Post-Quantum Cryptography 2019. LNCS, vol. 11505, pp. 167–186. Springer, Chongqing, China (May 2019). https://doi.org/10.1007/978-3-030-25510-7_10, https://doi.org/10.1007/978-3-030-25510-7_10
39. Wiedemann, D.: Solving sparse linear equations over finite fields. IEEE transactions on information theory **32**(1), 54–62 (1986)

**Proceedings of the ASME 2020 International Design Engineering Technical Conference &
Computers and Information in Engineering Conference
IMECE 202
August 16-19, 2020, St. Louis, USA**

IDETC2020-22137

ERROR QUANTIFICATION IN DYNAMIC APPLICATIONS OF WEAKLY NONLINEAR TRANSDUCERS

Lautaro Cilenti *
Department of Mechanical Engineering
University of Maryland - College Park
College Park, Maryland 20742
Email: lcilenti@terpmail.umd.edu

Akobuije Chijioke
Mass and Force Group
National Institute of Standards and Technology
Gathersburg, Maryland 20878
Email: Ako.Chijioke@nist.gov

Nicholas Vlajic
Applied Research Laboratory
The Pennsylvania State University
University Park, Pennsylvania 16802
Email: nav5000@arl.psu.edu

Balakumar Balachandran
Department of Mechanical Engineering
University of Maryland - College Park
College Park, Maryland 20742
Email: balab@umd.edu

ABSTRACT

Characterization and quantification of dynamic measurements is an ongoing area of research in the metrological community, as new calibration methods are being developed to address dynamic measurement applications. In the development undertaken to date, one largely assumes that nominally linear transducers can be used with linear assumptions in deconvolution of the input from the response and in system identification. To quantify the errors that arise from these assumptions, in this article, the effects of weak nonlinearities in transducers that are assumed to behave linearly during dynamic excitations are studied. Specifically, a set of first-order and second-order systems, which can model many transducers with weak nonlinearities, are used to numerically quantify the systemic errors due to the linear assumptions underlying the deconvolution. We show through the presented results the evolution of different error metrics over a large parameter space of possible transducers. Additionally, an example of quantification of the errors due to linear assumptions in system identification is demonstrated by using a time-series sparse regression system identification strategy. It is shown that the errors generated from linear identification of a nonlinear transducer can counteract the systemic errors that arise in linear deconvolution when the linear system identification is performed in similar loading conditions. In general, the methodology and results presented here can be useful for understanding the effect of nonlinearity in single degree of freedom transient dynamics deconvolution and specifically in specifying certain metrics of errors in transducers with known weak nonlinearities.

INTRODUCTION

Dynamic calibrations and signal deconvolution for transducers are a major area of development in the metrology community over the past decade [1, 2]. The developed framework for evaluating measurement uncertainty relies on the principle of superposition [3]. However, this principle does not apply for finding a response of a nonlinear system. Hence, weak nonlinearities in transducers are a source of unaccounted error.

Much work has been done to analyze the effects of weak nonlinear terms on the system response when the system is subjected to harmonic forcing. As an example, it is mentioned that Nayfeh and Mook [4] have shown how the frequency response of

[*] Address all correspondence to this author.

Cilenti, Lautaro; Chijioke, Akobuije; Balachandran, Balakumar. "Error Quantification in Dynamic Applications of Weakly Nonlinear Transducers." Presented at ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 2: 16th International Conference on Multibody Systems, Nonlinear Dynamics, and Control (MSNDC), St Louis, MO, US. August 16, 2020 - August 19, 2020.

a system with weak cubic nonlinearity can be approximated near a primary resonance, a sub-harmonic resonance, and a super-harmonic resonance by using perturbation methods. In general, these methods are meant for characterizing the responses of nonlinear systems subjected to a persistent excitation rather than a transient excitation such as an impulse. There are methods such as Volterra series based approaches, which have been used for determining the system response when subjected to an impulse input. Volterra series is a generalization of the convolution integral for nonlinear systems, and the Generalized Frequency Response Function (GFRF) may be considered as the equivalent one in the frequency domain [5].

A Volterra Series and GFRF consist of an infinite number of Volterra kernels; in practice they require truncation for actual application and can be computationally expensive for even low-order approximations [6, 7]. There has been previous work done on defining the generalized impulse response function for Gaussian impulses in applications related to the dynamic behavior of the economy [8]. In part due to the challenges encountered in identifying the effect of nonlinearity under arbitrary loading conditions, many transducers are designed to operate as linearly as possible, and in some cases, these transducers are subjected to nonlinear dynamics compensation strategies [9, 10].

Despite the efforts undertaken to date, transducers generally have some degree of nonlinearity in the responses which can lead to measurement errors. In this context, it is mentioned that we have found no previous work on quantification of the errors that arise in deconvolution when one assumes the responses of weakly nonlinear systems subjected to impulse loading conditions to be linear. Here, we use numerical methods to analyze a parameter space that we posit to be applicable to real transducers, and may be useful in quantifying the errors that arise due to the neglect of nonlinearities in nominally linear transducers.

In order to isolate the uncertainty due to the nonlinearity, other sources of potential measurement error are neglected. We assume noiseless measurements and that the dynamics of the system is exactly described by a first-order or second-order system, as shown in equation (1) and (2), respectively [11]. We focus primarily on transducers that are modeled as first-order systems, which can also be a suitable model for filtered certain second-order and higher systems. Some representative second-order system results are also included.

$$(a+\varepsilon_a)\frac{\mathrm{d}y}{\mathrm{d}t}+(\beta+\varepsilon_\beta)y+g(y)=F(t) \tag{1}$$

$$(a+\varepsilon_a)\frac{\mathrm{d}^2y}{\mathrm{d}t^2}+(\mu+\varepsilon_\mu)\frac{\mathrm{d}y}{\mathrm{d}t}+(\beta+\varepsilon_\beta)y+g(y,\frac{\mathrm{d}y}{\mathrm{d}t})=F(t) \tag{2}$$

In these equations, y represents the state of the systems and $F(t)$ represents the input. Errors in acquiring the parameters from imperfect system identification models are represented by the values with ε in the coefficients. The function $g(y)$ is used to represent the nonlinear terms. Systemic errors from possible additional linear terms are not shown here.

The coefficients in the nonlinear term $g(y)$ are assumed to be small, in keeping with the focus of this study on transducers that have been designed to exhibit near linear behavior. These equations are used to model the behavior of a nonlinear sensor and approximate the error in an inferred (deconvolved) input.

The rest of the paper is organized as follows: The analytical methodology that is used as a basis for the simulations is described in Section 2. The numerical methodology including useful metrics for quantifying time series errors in inferring (or deconvolving) the response inputs are presented in Section 3. Results from the numerical simulations are presented in Section 4. Finally, we present conclusions and remarks to close the article.

ANALYTICAL METHODOLOGY

First-Order Systems

Nondimensionalization In reducing the relevant parameter space for numerical study, the system can be represented in nondimensionalized form with the change of variables shown in equation (3). After using x as the new state variable and τ as the new time scale, equation (1) is transformed into equation (4).

$$t=\frac{a}{\beta}\tau,\quad y=\frac{1}{\beta}x \tag{3}$$

$$\frac{\mathrm{d}x}{\mathrm{d}\tau}+x+\tilde{g}(x)=\tilde{F}(\tau) \tag{4}$$

Equation (4) is the system in nondimensionalized form, where x represents the nondimensionalized state variable, and τ represents the nondimensionalized time. The function $\tilde{g}(x)$ represents the nondimensionalized nonlinear term, and $\tilde{F}(\tau)$ is an arbitrary input resulting from the nondimensionalization of x and τ. It was assumed in this nondimensionalization that ε is equal to zero; that is, the system parameters are known accurately.

Model-Based Linear Deconvolution We consider sensors that are accurately described by equation (4), and define a model-based linear deconvolution as evaluating the input from the output by using a linear differential equation with the same parameters as the nonlinear system.

$$\frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau}+\hat{x}+\tilde{g}(\hat{x})=\hat{\tilde{F}}(\tau) \tag{5}$$

$$F_{\text{estimated}} = \frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau} + \hat{x} \tag{6}$$

From Equation (5), it follows that for the input $\hat{\hat{F}}(\tau)$, the system has the response or solution $\hat{x}$. The error in the estimated input can be quantified by the difference between the estimated input and the actual input.

$$F_{\text{actual}} \equiv \hat{F}(\tau) \tag{7}$$

$$e(\tau) \equiv |F_{\text{estimated}} - F_{\text{actual}}| = |-\tilde{g}(\hat{x})| \tag{8}$$

Modeling Nonlinearity The systems described in equation (1) can be used to represent any arbitrary nonlinear system whose form depends on a transducer's behavior. In this paper, we restrict attention to the nonlinearities represented by equation (9).

$$\tilde{g}(x) = \alpha_2|x|x + \alpha_3 x^3 \tag{9}$$

The considered nonlinear terms are limited up to third order, as transducers are generally meant to be as linear as possible. The higher powers of nonlinearity are assumed to be small. This general function $g(x)$ will be adapted as necessary by setting the coefficients of nonlinearities that are not present in a particular system to zero.

Additional Error From System Identification It is of practical interest to consider the errors that arise when linear system identification is used to construct a linear model of a nonlinear transducer. So far it has been assumed that the parameter ε is zero, meaning the parameters are the same in the linear model as in the true nonlinear system. This assumption is valid when nonlinear system identification has been used to generate the parameters for a system described by these models and where the system identification measurement errors are small. In the practical scenario wherein the system model is constructed from linear system identification, ε will include a component due to the ignored or unidentified nonlinearity during the system identification. Neglecting other sources of errors, we assign the value ε_o to ε. The parameter ε_o is zero when using nonlinear system identification for a nonlinear system and nonzero in general when using linear system identification for a nonlinear system.

When ε_o is nonzero, equation (8) is only valid approximately. This is further developed as follows: The actual nonlinear system has ε equal to zero and is described in nondimensionalized form by equation (4).

Alluding here to equation (1) which includes errors in the identified system parameters, in an application where the user has assumed the system to be linear and performed a noise free linear system identification, the corresponding model has nonzero $\varepsilon = \varepsilon_o$. This linear model is described by equation (10), where $\rho_n = \varepsilon_{n,o}/n$.

$$(1+\rho_a)a\frac{\mathrm{d}y}{\mathrm{d}t} + (1+\rho_\beta)\beta y = F(t) \tag{10}$$

In nondimensionalization of the linear system, the change of variables also takes the form shown in equation (3). The corresponding nondimensionalized versions of equation (10) is shown as equation (11).

$$(1+\rho_a)\frac{\mathrm{d}x}{\mathrm{d}\tau} + (1+\rho_\beta)x = \tilde{F}(\tau) \tag{11}$$

After substituting the particular solution $\hat{x}$ to obtain $F_{\text{estimated}}$ in equation (11) and taking the difference with F_{actual}, one obtains $e(\tau)$. Taking into account the ε_o error, equation (8) is replaced by equation (12).

$$e(\tau) \equiv |F_{\text{estimated}} - F_{\text{actual}}| = |-\tilde{g}(\hat{x}) + \xi(\tau)| \tag{12}$$

$$\xi(\tau) = \rho_a\frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau} + \rho_\beta\hat{x} \tag{13}$$

The case where all $\rho = 0$ corresponds to the case where there is no nonlinearity. Also $\rho = 0$ for the case of perfect nonlinear system identification, and in which case equation (12) is equivalent to equation (8).

In this analysis, the type of nonlinearity or magnitude of the nonlinearity itself has not been considered. In a practical application, the type and magnitude of nonlinearity will determine the magnitudes of ρ.

Magnitude of System Identification Error, ρ At this point, it is unclear as to how large ρ can be for a particular nonlinearity. Since it is possible that (13) will be of the same magnitude as $\tilde{g}(\hat{x})$ for weak nonlinearities, it is necessary to understand how ρ changes with the nonlinearity to understand the applicability of this methodology in applications where nonlinear system identification is not performed. To quantify the magnitude of ρ for a particular nonlinearity and forcing amplitude, we selected a system identification methodology based on Kutz and colleagues' nonlinear identification of governing equations using sparse regression [12]. This methodology is meant to serve only as an example, as it is likely that other linear system identification strategies will result in different errors in the parameters.

In nonlinear identification carried out with the method of Kutz and colleagues, a simulated nonlinear output, $\hat{x}$ from equation (4), is used to generate a matrix A of candidate nonlinear functions as shown in equation (14). Each row of the matrix corresponds to a time step of the discrete $\hat{x}$ and each column is a nonlinear transformation of $\hat{x}$ in terms that are candidates for the nonlinear model.

$$A_{nl} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ \hat{x} & |\hat{x}|\hat{x} & \hat{x}^3 & \hat{\tilde{F}} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{14}$$

We use this methodology while ignoring the nonlinear terms, allowing the regression to optimize the linear parameters to the nonlinear trajectory $\hat{x}$. A matrix is generated with only the terms in the linear differential equation as shown in equation (15).

$$A_l = \begin{pmatrix} \vdots & \vdots \\ \hat{x} & \hat{\tilde{F}} \\ \vdots & \vdots \end{pmatrix} \tag{15}$$

$$A_l \omega_l = \hat{x}_{dot} \tag{16}$$

In equation (16), $\hat{x}_{dot}$ is the time derivative of $\hat{x}$ computed by using the central difference formula, and ω_l is a vector of weights optimized for the governing equation that generated $\hat{x}$. After solving equation (16) with a solver that promotes sparsity, one obtains weights, ω_l, and generates equation (17).

Equation (18) is equivalent to (11). Comparing equation (17) to equation (18), we define equations (19) and (20).

$$\hat{x}_{dot} - \omega_l(1)\hat{x} = \omega_l(2)\hat{\tilde{F}} \tag{17}$$

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} + \frac{(1+\rho_\beta)}{(1+\rho_a)}x = \frac{1}{(1+\rho_a)}\tilde{F} \tag{18}$$

$$\omega_l(1) \equiv \frac{-(1+\rho_\beta)}{(1+\rho_a)} \tag{19}$$

$$\omega_l(2) \equiv \frac{1}{(1+\rho_a)} \tag{20}$$

From equations (19) and (20), we derive ρ_a and ρ_b as shown in equation (21).

$$\rho_a = \frac{1}{\omega_l(2)} - 1, \quad \rho_b = -\omega_l(1) \times (1+\rho_a) - 1 \tag{21}$$

The ρ values obtained following this methodology can be used to quantify $e(\tau)$ as shown in equation (12).

Second-Order Systems

Here, we consider the second-order systems described by equation (2). The analysis described in this section follows along the same lines as that presented for the first-order systems and will be presented in condensed form.

The nondimensional version of the second-order system uses the change of variables in equation (22) to generate equation (23) from equation (2).

$$t = \sqrt{\frac{a}{\beta}}\tau, \quad y = \frac{1}{\beta}x, \quad \delta = \frac{\mu}{\sqrt{a\beta}} \tag{22}$$

$$\frac{\mathrm{d}^2x}{\mathrm{d}\tau^2} + \delta\frac{\mathrm{d}x}{\mathrm{d}\tau} + x + \tilde{g}(x, \frac{\mathrm{d}x}{\mathrm{d}\tau}) = \tilde{F}(\tau), \tag{23}$$

The parameter δ represents the nondimensionalized linear damping coefficient. The nonlinear terms in the second-order systems are described with $\tilde{g}(x, \frac{\mathrm{d}x}{\mathrm{d}\tau})$ that includes the lower order nonlinear velocity terms as shown in equation (24).

$$\tilde{g}(x, \frac{\mathrm{d}x}{\mathrm{d}\tau}) = \alpha_2|x|x + \alpha_3x^3 + \gamma_1 x\frac{\mathrm{d}x}{\mathrm{d}\tau} + \gamma_2\frac{\mathrm{d}x}{\mathrm{d}\tau}\left|\frac{\mathrm{d}x}{\mathrm{d}\tau}\right| \tag{24}$$

Following a similar approach as that for the first-order systems, the error, $e(\tau)$, can be calculated by taking the difference between an actual input and an estimated input calculated by applying the output of the nonlinear system, $\hat{x}$, to the linear system shown in equation (26).

$$\frac{\mathrm{d}^2\hat{x}}{\mathrm{d}\tau^2} + \delta\frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau} + \hat{x} + \tilde{g}(\hat{x}, \frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau}) = \hat{\tilde{F}}(\tau) \tag{25}$$

$$F_{\text{estimated}_2} = \frac{\mathrm{d}^2\hat{x}}{\mathrm{d}\tau^2} + \delta\frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau} + \hat{x} \tag{26}$$

$$F_{\text{actual}} \equiv \hat{F}(\tau) \tag{27}$$

$$e(\tau) \equiv \left|F_{\text{estimated}_2} - F_{\text{actual}}\right| = \left|-\tilde{g}(\hat{x}, \frac{\mathrm{d}\hat{x}}{\mathrm{d}\tau})\right| \tag{28}$$

Note that the definition of $e(\tau)$ is in the same spirit for the second-order systems as that for the first-order systems.

Errors caused by nonlinearity in system identification can be quantified in a similar manner for the second-order systems as for the first order systems, but we do not do so here.

SIMULATION METHODOLOGY

Numerical Methods

Numerical integration of differential equations (4) and (23) was performed by using the Dormand-Prince method with a relative tolerance of 10^{-8} and an absolute tolerance of 10^{-8}. In order to obtain estimated F, the derivatives of the solutions were calculated by using a central difference formula. Since, in the Dormand-Prince method, one does not attempt to minimize errors in the discrete derivative (relevant especially when discontinuities are present), the simulations were forced to run with step sizes smaller than or equal to h_{max}. The value of h_{max} was fixed to 0.01 based on visual inspection of convergence while minimizing simulation run-time. The simulations ran with different values of the parameters α_2, α_3, γ_1, and γ_2 for the nonlinearities given in equations (9) and (24). Only one nonlinear term was nonzero in a given simulation; combinations of nonlinearities were not studied. The input, $\tilde{F}(\tau)$, was a Gaussian envelope impulse described by equation (29) generated from the SciPy Python library [13], in which amplitude A and duration $\Delta\tau$ were varied.

$$\tilde{F}(\tau) = Ae^{C_1\tau^2}, \quad C_1 = -\frac{14.288}{(\Delta\tau)^2} \tag{29}$$

For finding the parameter space that generates a prescribed error due to nonlinearity, we ran the simulations by using an iterative solver based on Newton's method to find the input amplitude, A, that generates a prescribed error for a fixed magnitude of the nonlinear coefficient. Newton's method was automated to stop at convergence with thresholds of at most 10 % the magnitude of the prescribed error.

The open source Scikit-learn lasso regression module [14] available on python was used to perform sparse regression in implementing Kutz's and colleagues' system identification methodology. A small regularizer in lasso regression promotes sparsity. The regularizing parameter, λ, was fixed to 10^{-6}. The regression module was also configured to run a maximum of 10^5 iterations or to stop at convergence tolerances of 10^{-6}. When performing nonlinear system identification as a control, this configuration kept the errors in the weights of the system identification methodology at around 0.1 % when used with simulated input amplitudes greater than 10 and smaller than 10^4. The additional systemic error generated from ignoring nonlinear terms in the linear identification regression are reported in the results.

Error Metrics

Since the response of the system and its errors are a function of time for dynamic loading, some useful metrics for dynamic error quantification are presented within this section. When the maximum error due to nonlinearity is of interest, the maximum point by point error can be quantified by using equation (30).

$$M_1 \equiv e_{\text{max}} = \frac{max(|e(\tau)|)}{max(|F_{\text{estimated}}(\tau)|)} \tag{30}$$

When the integrated input is of interest, the relevant errors can be quantified by using equation (31).

$$M_2 \equiv e_{\text{integral}} = \frac{\int_0^{\tau_f} e(\tau)d\tau}{\int_0^{\tau_f} F_{\text{estimated}}(\tau)d\tau} \tag{31}$$

In equation (31), the denominator represents the integral of estimated F with respect to time and the numerator represents the integral of the error in estimated F due to nonlinearity with respect to time. The parameter $\tau_f >> \Delta\tau$, meaning that the integration continues until the transient response due to an input of duration $\Delta\tau$ has died out.

When the peak input is of interest, the error can be quantified by equations (32)-(34).

$$\tau_{\text{max1}} = \text{argmax}(|F_{\text{actual}}(\tau)|) \tag{32}$$

$$\tau_{\text{max2}} = \text{argmax}(|F_{\text{estimated}}(\tau)|) \tag{33}$$

$$M_3 = e_{\text{PeakRatio}} = \frac{|F_{\text{actual}}(\tau_{\text{max1}}) - F_{\text{estimated}}(\tau_{\text{max2}})|}{|F_{\text{estimated}}(\tau_{\text{max2}})|} \tag{34}$$

The denominator in equation (34) represents the peak estimated input, and the numerator represents the difference in peaks between the actual input and the estimated input.

The metrics are applied in post-processing of simulations in parameter space of input amplitude, A, input duration, $\Delta\tau$, and nonlinearity coefficients from equations (9) and (24). They can be used to obtain uncertainty estimates when an inferred peak or integrated input is known by multiplying the error ratios with the experimental inferred input values.

RESULTS

First-Order System Results

Simulations of first-order systems with quadratic or cubic nonlinearity subjected to a gaussian impulse were conducted. From the response, an inferred linearly deconvolved input was calculated by neglecting the nonlinear term. An inferred input from a nonlinear system with a relatively large cubic nonlinear term compared to the actual input is shown in Fig. 1. The three error metrics are evaluated in post-processing. From Fig. 2, one can observe the parameter space of forcing amplitude and nonlinearity coefficient where the errors at the peaks (equation (34)) are 10 % the inferred input amplitude. We note that the cubic nonlinearity has a slope of $-\frac{1}{2}$ and the quadratic nonlinearity has a slope of -1 in the log space of input amplitude A and nonlinearity coefficient. These slopes are consistent for the three different error metrics and for different $\Delta\tau$. Lines corresponding to error at the peak (Metric 3: equation (34)) of 1 %, 10 %, 20 %, 50 %, and 90 % for a cubic nonlinearity are shown in Fig. 3. It can be inferred from this graph that transducers that are used with input amplitudes smaller than the 1 % line yield error metric values smaller than 1 % when used to infer an input. We show in Fig. 4 the vertical shift in the lines for the three different error metrics for a cubic nonlinearity with a fixed metric value of 10 %. Metric 1, representing the maximum error between the inferred input and the actual input (equation (30)) yields 10 % errors at lower input amplitudes than metric 2, the error at the peaks (equation (34)). It can be inferred from this observation that the maximum error in deconvolution does not occur at the inferred peak; this

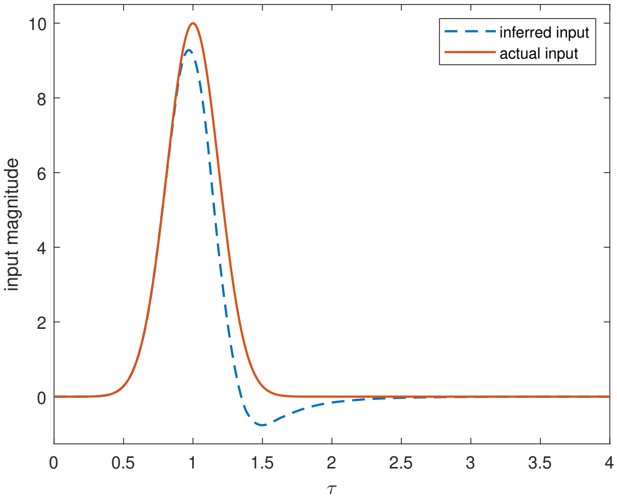


FIGURE 1. Inferred input compared to the actual input for a nonlinear system with $\tilde{g}(x) = 0.1x^3$

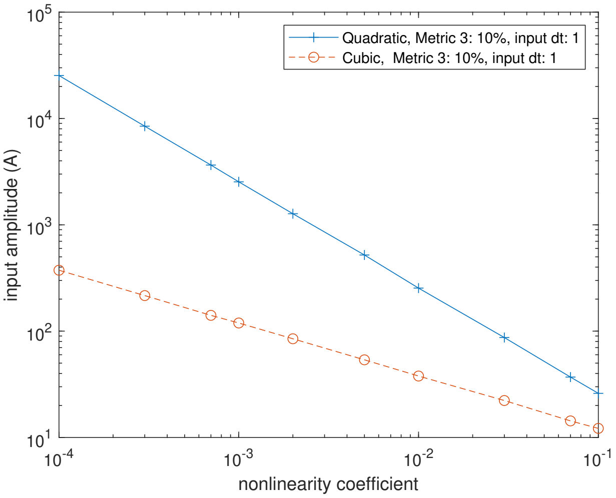


FIGURE 2. Gaussian input amplitude plotted with nonlinearity coefficient for quadratic ($|x|x$) and cubic (x^3) nonlinear terms that produce error at peak metric values of 10 %. The errors are smaller than 10 % in the parameter space below the lines.

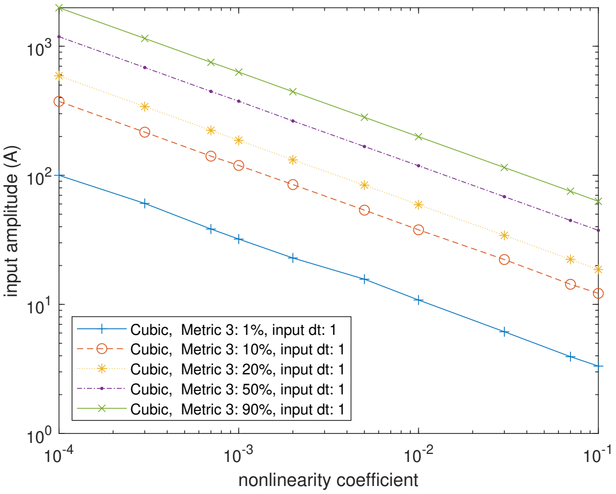


FIGURE 3. Gaussian input amplitude plotted with nonlinearity coefficient for cubic (x^3) nonlinear terms that produce errors of varying magnitude.

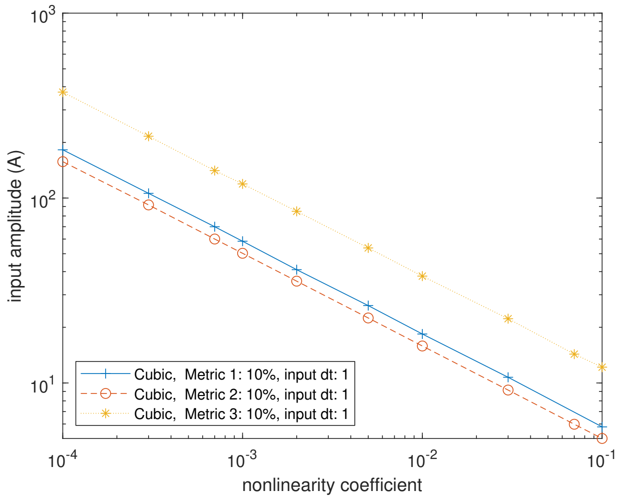


FIGURE 4. Gaussian input amplitude plotted with nonlinearity coefficient for cubic (x^3) nonlinear terms that produce error magnitudes of 10 % for the three different error metrics.

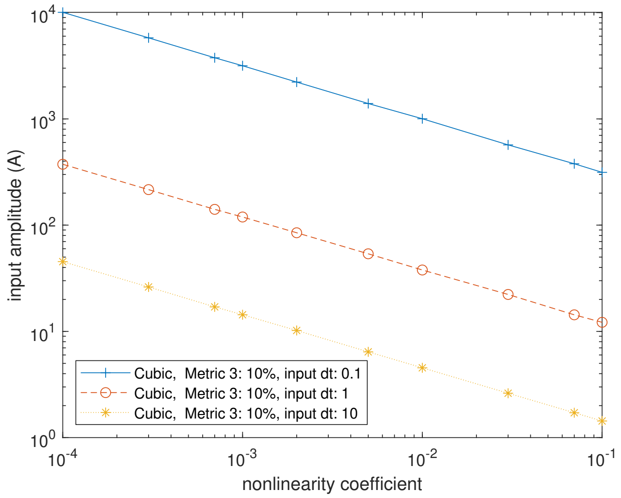


FIGURE 5. Gaussian input amplitude plotted with nonlinearity coefficient for cubic (x^3) nonlinear terms that produce error magnitudes of 10 % for metric 3 with three different input durations.

TABLE 1. Parameters m_1, m_2, and b can be used in equation (35) to approximate the error value for a particular system. Here, 1^* is bound as $1 > 1^* \geq 0.995$.

Δt	Metric	x^3				$\|x\|x$			
		m_1	m_2	b	R^2	m_1	m_2	b	R^2
0.1	1	-0.5	0.49	2.21	0.90	-1	1.40	3.08	1*
0.1	2	-0.5	0.91	2.17	0.95	-1	2.37	2.97	0.98
0.1	3	-0.5	0.68	2.71	0.99	-1	1.22	3.51	1*
1	1	-0.5	0.73	0.99	1*	-1	1.30	1.27	1*
1	2	-0.5	0.68	0.92	0.99	-1	1.48	1.22	0.99
1	3	-0.5	0.64	1.2	0.99	-1	1.20	1.6	1*
10	1	-0.5	0.61	0.26	0.99	-1	1.22	0.33	1*
10	2	-0.5	0.65	0.43	0.98	-1	1.26	0.49	1*
10	3	-0.5	0.6	0.29	0.99	-1	1.18	0.37	1*

can also be observed in the example of Fig. 1. The vertical shift in the lines for three different $\Delta\tau$ with a fixed metric value of 10 % are shown in Fig. 5. We note that $\Delta\tau = 0.1$ corresponds to the largest bandwidth of excitation in this study. It is expected that the Metric 1 and 3 lines will converge to the static case at the limit $\Delta\tau \to \infty$. From observations made in figs. 2 to 5, we generated Table 1 and equation (35) as an interpolating model for the included figures and those not included. This can be used to generate an approximate error metric value corresponding to a nonlinear coefficient, α, and impulse amplitude, A.

$$M(\alpha,A) = 100 \times 10^{\left(\frac{\log_{10}(A) - m_1\log_{10}(\alpha) - b}{m_2}\right)} \tag{35}$$

As shown in Table 1, the parameters m_1, m_2, and b have R^2 values close to 1. This indicates that the linear regression in the interpolating model fit the data well. In the parameter range shown in the included figures, the interpolating model is accurate within a factor of 2.

First-Order System Identification Error

After using Kutz's and colleagues' system identification from data approach, $e(\tau)$ was recalculated considering the error in linear system identification of a nonlinear system. The new $e(\tau)$ from equation (12) produces new metric values and shifts metric lines as shown in Fig. 6. One can observe from Fig. 6 that the metric lines of 10 % that include nonzero ρ error occur at higher forcing amplitudes than the original system. One way to understand this behavior is to note that the resulting sign of ρ has the effect of counteracting the negative $\tilde{g}(\hat{x},\tau)$ in $e(\tau)$; a smaller $e(\tau)$ yields a smaller metric value resulting in higher input amplitude or nonlinear coefficients to produce the same error. A more intuitive way of understanding the effect is to observe that the optimization of the linear parameters to the nonlinear data by using linear system identification reduced the error between the linear inferred input and the actual input because it was optimized to do so through an adjustment in the linear parameters during regression. As shown in Fig. 7, the same trend occurs for all three metrics at 10 %. We note at this point that the ρ values were obtained with linear system identification at same input amplitude that was used for the inferring procedure. Unlike in the

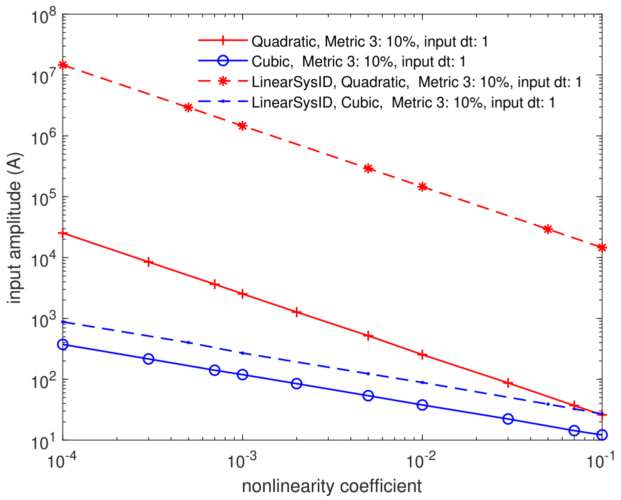


FIGURE 6. Gaussian input amplitude plotted with nonlinearity coefficient for quadratic ($|x|x$) and cubic (x^3) nonlinear terms that produce error at peak metric values of 10 %. The plot compares the actual nonlinear system results with the results of considering the ρ_a and ρ_b errors generated from performing linear system identification on the nonlinear system.

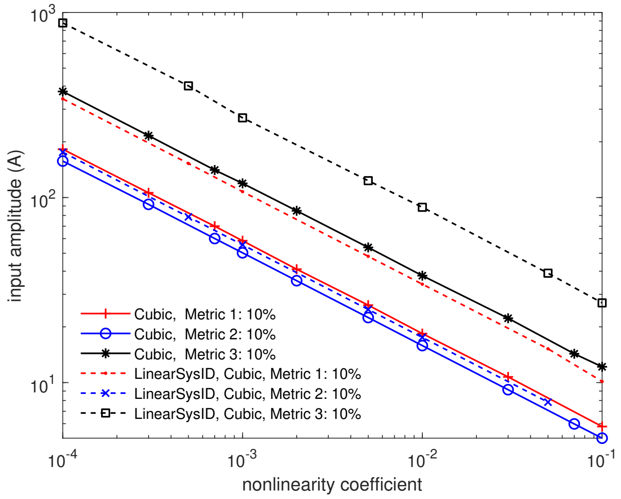


FIGURE 7. Gaussian input amplitude plotted with nonlinearity coefficient for cubic (x^3) nonlinear term that produce error at peak metric values of 10 %. The plot compares the actual nonlinear system results with the results of considering the ρ_a and ρ_b errors generated from performing linear system identification on the nonlinear system for the three metrics.

case of nonlinear system identification of the parameters, the optimization of the parameters by using linear system identification is only valid at the amplitude of the input used for generating $\hat{x}$ prior to regression. With this in mind, we show the contour plots in Fig. 8 and 9 which illustrate how ρ changes in the parameter space.

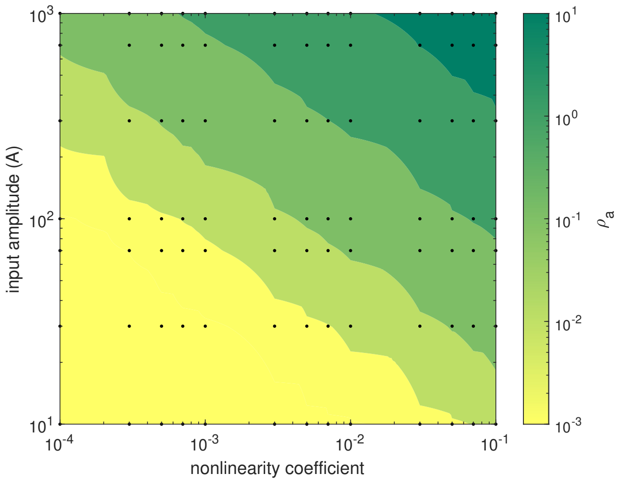


FIGURE 8. ρ_a magnitude in parameter space of input amplitude and nonlinearity coefficient for system with cubic nonlinearity. Simulated points are shown as black dots.

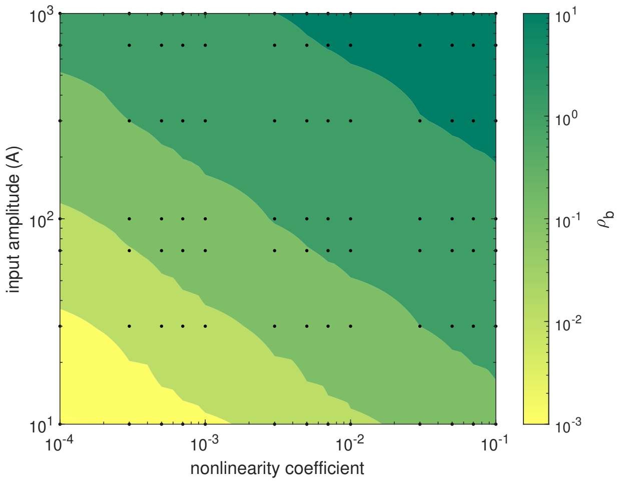


FIGURE 9. ρ_b magnitude in parameter space of input amplitude and nonlinearity coefficient for system with cubic nonlinearity. Simulated points are shown as black dots.

The parameter space shown in Fig. 8 and 9 contains only positive values of ρ. For larger input amplitudes and lower input amplitudes than presented, the sparse regression failed to converge. This is likely a direct result of the tight tolerances and small regularizer used for lasso regression, but may be related to the ability to identify a nonlinear system through linear identification when excited by very small amplitudes or very large amplitudes.

Second-Order System Results

From the responses of second-order systems with nonlinearities subjected to a Gaussian impulse an inferred linearly deconvolved input was calculated by neglecting the nonlinear term. The inferred input for four different types of nonlinearities are shown in Fig. 10.

From Fig. 10 and explorations of the parameter space for the second-order system, it can be qualitatively observed that the nonlinearity does not have a large effect on the peak inferred input. However, the nonlinearity modulates the frequency of the response and generates transient oscillations in the inferred input after the first peak that decay at rates depending on the linear or nonlinear damping. Furthermore, as shown in Fig. 11, for higher forcing amplitudes of the same nonlinear systems, the peak inferred inputs can occur at the second peak of oscillation. This indicates strongly nonlinear behavior and transducers that exhibit this behavior would not be considered nominally linear.

Based on these qualitative observations, we note that Metric 3 which quantifies errors at the peaks is of small magnitude for the second order system until there is such strongly nonlinear behavior that the second peak can be larger than the first peak. Furthermore, due to the contribution of transient oscillations after the first peak in highly nonlinear behavior, the error metrics can be highly dependent on the damping ratio of the system.

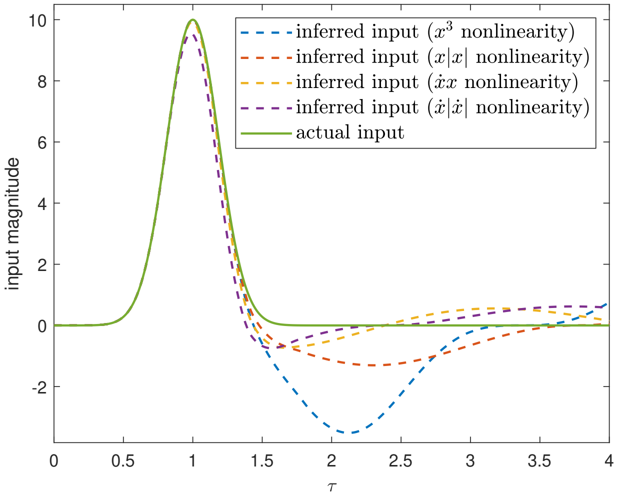


FIGURE 10. Inferred inputs compared to the actual input for nonlinear second order systems with $\tilde{g}(x, \frac{\mathrm{d}x}{\mathrm{d}\tau})$ of $0.1x^3$, $0.1x|x|$, $0.1x\frac{\mathrm{d}x}{\mathrm{d}\tau}$, and $0.1\frac{\mathrm{d}x}{\mathrm{d}\tau}\left|\frac{\mathrm{d}x}{\mathrm{d}\tau}\right|$ and damping term $\delta = 0.2$.

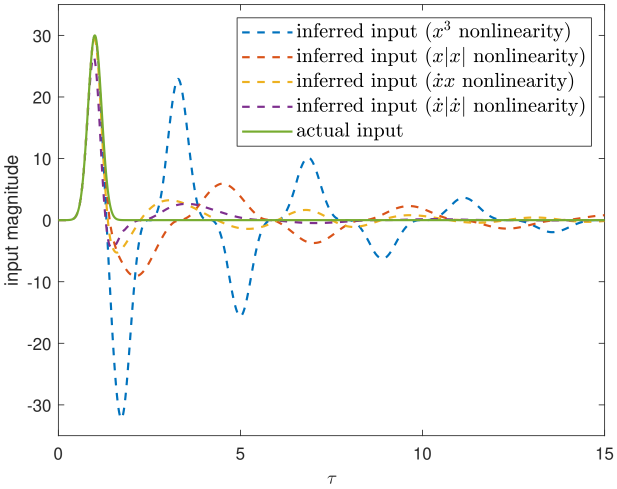


FIGURE 11. Inferred inputs compared to the actual input for nonlinear second order systems with $\tilde{g}(x, \frac{\mathrm{d}x}{\mathrm{d}\tau})$ of $0.1x^3$, $0.1x|x|$, $0.1x\frac{\mathrm{d}x}{\mathrm{d}\tau}$, and $0.1\frac{\mathrm{d}x}{\mathrm{d}\tau}\left|\frac{\mathrm{d}x}{\mathrm{d}\tau}\right|$ and damping term $\delta = 0.2$. Notice the second peak can be higher than the first for the inferred input.

We limit the scope of the analysis of the second order system in this paper to a few examples. From Fig. 12, one can observe the parameter space by which errors at the peaks remain below 1 ± 0.1 % for the four different nonlinear terms. As shown in Fig. 13 for systems with cubic nonlinearity the error rises quickly after it has reached 1 %. This occurs because at the 1 % line for the cubic nonlinearity, the second peak is larger than the first peak and it continues to grow with increasing input amplitudes. One may take away from this nonetheless that certain nonlinearities will produce errors smaller that 1 % up to the point they are so large that they are producing high second peaks in the inferred input.

We show in Fig. 14 how Metric 2, the error in the integral of the inferred input, decreases with increasing damping ratio for a system with the nonlinear function $g(x, \frac{\mathrm{d}x}{\mathrm{d}\tau}) = 0.01x^3$. This behavior is also observed for the other three nonlinear terms and also generates lower Metric 1 and Metric 2 errors with greater damping as it takes higher input amplitudes and nonlinear coefficients to generate high oscillations after the first inferred peak.

Finally we also show an example in Fig. 15 where changes to the excitation duration for the second order system shifts the error lines vertically as in the first order system. This is most likely due to the corresponding change in excitation momentum when the excitation amplitudes are held constant but the excitation duration is changed.

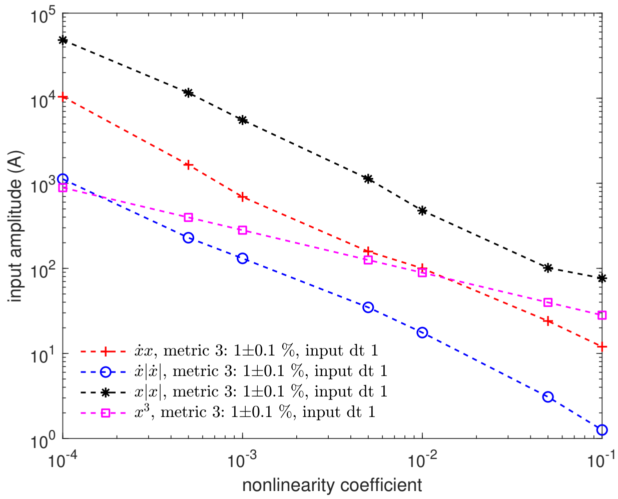


FIGURE 12. Error at the peaks of 1 ± 0.1 % between inferred input and actual input for four different nonlinear terms in a second-order systems with damping ratio $\delta = 0.2$ excited by a one second pulse.

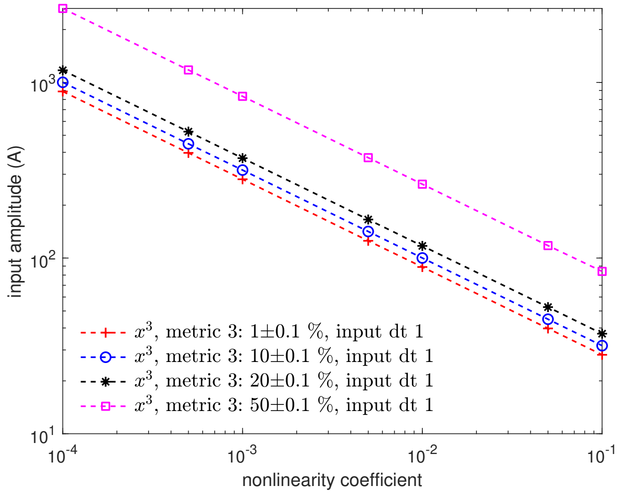


FIGURE 13. Error at the peaks between inferred input and actual input for cubic nonlinearity in second-order systems with damping ratio $\delta = 0.2$ excited by a one second pulse.

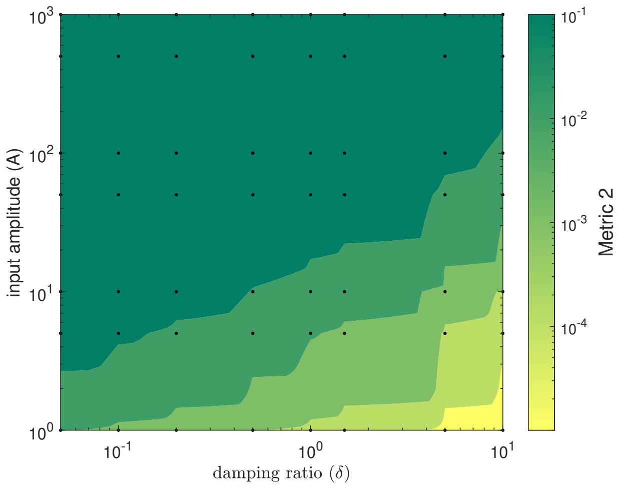


FIGURE 14. Filled contour of integral error ratio (metric 2) as a function of damping ratio and input amplitude for a fixed nonlinearity, $0.01x^3$.

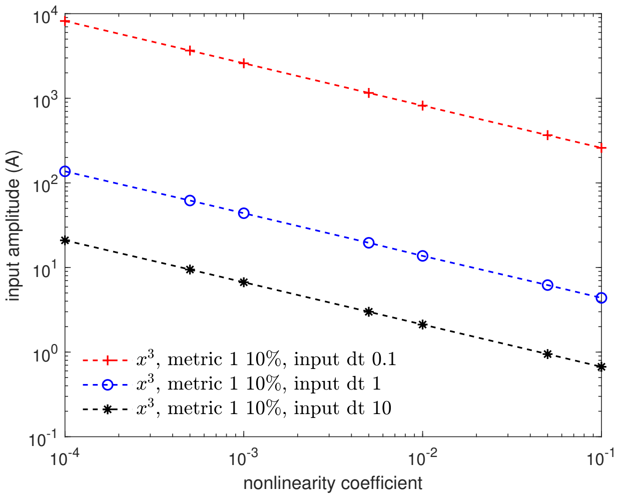


FIGURE 15. Maximum error between inferred input and actual input for three different bandwidths of excitations of a second order system with cubic nonlinearity and damping ratio $\delta = 0.2$.

DISCUSSION

A novel methodology for quantifying errors in the decovolution process in transducers, which can be modeled as first- or second-order systems is presented within the article. This methodology is not limited to nominally linear transducers. Here, the lower order nonlinear terms more prevalent in weakly nonlinear systems are considered. We also present error metrics that may be useful in quantifying errors in measurements with nominally linear transducers. Through simulations, we highlight the variations of the error metrics in a space of weak nonlinear coefficients. If upper bounds to the nonlinear coefficients and the input amplitudes for a transducer are known, then upper bounds to the error metrics can be estimated. Furthermore, we provide an interpolating function for estimating errors in deconvolution over the parameter space that was studied here.

We present a simple and novel methodology for quantifying the error that arises in ignoring the nonlinearity in a transducer during system identification. Through simulations, we show that using this approach, in first-order systems within the parameter space of this study, errors due to system identification can counter the errors in deconvolution by optimizing the inferred input to the actual input by using the linear parameters. A metrologist may be interested in this effect, noting certain linear calibration methods performed near measuring conditions for a nonlinear sensor may help diminish unaccounted errors in linear deconvolution.

For the second-order systems, we also present some examples where the errors in the second-order system are observed as transient oscillations after the first peak and decrease with higher linear damping ratios.

The parameter space studied is not comprehensive for nonlinear systems in general or nominally linear transducers. Additionally, only Gaussian impulses were explored and for only three limited bandwidths. Furthermore, the effects of noise were not considered in the methodology and simulations.

It would be interesting in future studies to compare this methodology and the associated results to results from deconvolution algorithms, to evaluate the extent to which real transducers fall within the parameter space considered here, and to test the methodology experimentally through real or representative transducers.

ACKNOWLEDGMENT

Partial support received for this work through the Non-Academic Research Internship for Graduate Students (INTERN) program provided by the U. S. National Science Foundation through Grant No. CMMI 1760366 is gratefully acknowledged.

Lautaro Cilenti is a Clark Doctoral Fellow of the Clark School of Engineering at the University of Maryland - College Park.

REFERENCES

[1] Eichstädt, S., Elster, C., Esward, T., and Hessling, P., 2010. “Deconvolution filters for the analysis of dynamic measurement processes: A tutorial”. *Metrologia,* ***47***, Oct., pp. 522–533.

[2] Eichstädt, S., Wilkens, V., Dienstfrey, A., Hale, P., Hughes, B., and Jarvis, C., 2016. “On challenges in the uncertainty evaluation for time-dependent measurements”. *Metrologia,* ***53***(4), June, pp. S125–S135. Publisher: IOP Publishing.

[3] Esward, T., Eichstädt, S., Smith, I., Bruns, T., Davis, P., and Harris, P., 2018. “Estimating dynamic mechanical quantities and their associated uncertainties: application guidance”. *Metrologia,* ***56***(1), Nov., p. 015002. Publisher: IOP Publishing.

[4] Nayfeh, A., and Mook, D., 2008. *Nonlinear Oscillations*. Wiley Classics Library. Wiley.

[5] Rijlaarsdam, D., Nuij, P., Schoukens, J., and Steinbuch, M., 2017. “A comparative overview of frequency domain methods for nonlinear systems”. *Mechatronics,* ***42***, Apr., pp. 11–24.

[6] Billings, S. A., and Peyton Jones, J. C., 1990. “Mapping non-linear integro-differential equations into the frequency domain”. *International Journal of Control,* ***52***(4), Oct., pp. 863–879.

[7] B. Zhang, S. A. Billings, Z. Lang, and G. R. Tomlinson, 2009. “Analytical Description of the Frequency Response Function of the Generalized Higher Order Duffing Oscillator Model”. *IEEE Transactions on Circuits and Systems I: Regular Papers,* ***56***(1), Jan., pp. 224–232.

[8] Potter, S. M., 2000. “Nonlinear impulse response functions”. *Journal of Economic Dynamics and Control,* ***24***(10), Sept., pp. 1425–1446.

[9] D. Yu, F. Liu, P. Lai, and A. Wu, 2008. “Nonlinear Dynamic Compensation of Sensors Using Inverse-Model-Based Neural Network”. *IEEE Transactions on Instrumentation and Measurement,* ***57***(10), Oct., pp. 2364–2376.

[10] Zimmerschied, R., and Isermann, R., 2010. “Nonlinear time constant estimation and dynamic compensation of temperature sensors”. *Control Engineering Practice,* ***18***(3), Mar., pp. 300–310.

[11] Vlajic, N., and Chijioke, A., 2015. “Modelling the Response of Force Transducers Under Sine-Sweep Calibration”. In IDETC-CIE2015. V008T13A055.

[12] Brunton, S. L., Proctor, J. L., and Kutz, J. N., 2016. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems”. *Proceedings of the National Academy of Sciences,* ***113***(15), Apr., p. 3932.

[13] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, l., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma,

M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., Vázquez-Baeza, Y., and SciPy 1.0 Contributors, 2020. "SciPy 1.0: fundamental algorithms for scientific computing in Python". *Nature Methods,* ***17***(3), Mar., pp. 261–272.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, d., 2011. "Scikit-Learn: Machine Learning in Python". *J. Mach. Learn. Res.,* ***12***(null), Nov., pp. 2825–2830. Publisher: JMLR.org.

Quasi-Deterministic Channel Model for mmWaves: Mathematical Formalization and Validation

Mattia Lecci*, Michele Polese‡, Chiehping Lai†, Jian Wang†, Camillo Gentile†, Nada Golmie†, Michele Zorzi*

*Department of Information Engineering, University of Padova, Italy, e-mail: {name.surname}@dei.unipd.it
‡Institute for the Wireless Internet of Things, Northeastern University, Boston, MA, USA, e-mail: m.polese@northeastern.edu
†National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, e-mail: {name.surname}@nist.gov

***Abstract*—5G and beyond networks will use, for the first time ever, the millimeter wave (mmWave) spectrum for mobile communications. Accurate performance evaluation is fundamental for the design of reliable mmWave networks, with accuracy rooted in the fidelity of the channel models. At mmWaves, the model must account for the spatial characteristics of propagation since networks will employ highly directional antennas to counter the much greater pathloss. In this regard, Quasi-Deterministic (QD) models are highly accurate channel models, which characterize the propagation in terms of clusters of multipath components, given by a reflected ray and multiple diffuse components of any given Computer Aided Design (CAD) scenario. This paper introduces a detailed mathematical formulation for QD models at mmWaves, that can be used as a reference for their implementation and development. Moreover, it compares channel instances obtained with an open source National Institute of Standards and Technology (NIST) QD model implementation against real measurements at 60 GHz, substantiating the accuracy of the model. Results show that, when comparing the proposed model and deterministic rays alone with a measurement campaign, the Kolmogorov-Smirnov (KS) test of the QD model improves by up to 0.537.**

***Index Terms*—5G, millimeter wave, channel model, 3GPP, IEEE, quasi deterministic**

I. Introduction

To satisfy a constantly growing demand for mobile connectivity, the future generations of wireless networks will exploit frequencies above 6 GHz for radio access. This portion of the spectrum, loosely identified as the millimeter wave (mmWave) band, features large chunks of untapped spectrum, to be used to provide ultra-high datarates to end users. In this regard, cellular networks implementing the 3GPP NR Release 15 specifications can support a carrier frequency of up to 52.6 GHz, while IEEE 802.11ad/ay foresee Wireless Local Area Networks (WLANs) operating in the unlicensed spectrum at 60 GHz. The development of robust mobile networks in this frequency range is challenging. The high propagation loss, indeed, limits the coverage of the mmWave base stations and access points. Besides, mmWave signals are blocked by common obstacles (e.g., the human body, walls, vehicles) with high penetration losses, making the power of the received signal highly variable.

A reliable and accurate evaluation of the performance is fundamental to the development of technological solutions for mmWave cellular networks. Given the difficulties associated to a testbed setup at such frequencies, the research community has, so far, mostly relied on analysis and simulations [1], developing several tools for different protocol stacks and proposed communication technologies [2]–[4]. The accuracy of the performance evaluation, however, depends to a large degree on the fidelity of the representation of the channel [5]. When it comes to mmWaves, the complex dynamics of the propagation environment strengthen the need for a comprehensive model, which accounts not only for the pathloss, but also for the spatial behavior of the signal propagation and its interaction with directional antennas, and for the fading that arises from the interaction with the scatterers in the environment [6].

This need has sparked several research efforts aimed at characterizing the mmWave channel. Measurement campaigns have been conducted in diverse settings, e.g., in urban or rural scenarios [7], [8], or indoors [9], [10]. These works have identified a number of key elements for the modeling of mmWave channels [11], [12]: (i) the multipath components are sparse in the angular domain, and this impacts the characterization and design of beamforming schemes; (ii) blockage affects the link dynamics much more than at sub-6 GHz; (iii) effects of diffuse scattering from rough surfaces become more prominent at shorter wavelengths. So far, different modeling approaches have emerged in the mmWave domain. The simplest ones are used, generally, for mathematical analysis, and characterize fading with Nakagami-m or Rayleigh random variables, often with simplified beamforming patterns [1]. The 3rd Generation Partnership Project (3GPP) has adopted a Spatial Channel Model (SCM) for the evaluation of NR in the frequency range between 0.5 and 100 GHz, in which a channel matrix is generated with a purely stochastic approach [13]. These approaches, however, cannot fully capture the fading and angular components of the mmWave channel that relate with a *realistic* and *specific* propagation environment.

This can be achieved using a Ray-Tracer (RT) [14], which models the channel by generating the Multi Path Components (MPCs) that, given the description of a certain scenario, can physically propagate from the transmitter's to the receiver's location. These MPCs are characterized by angles of arrival and departure, power and delay, and can either be the direct component, or rays reflected from the scattering surfaces of the environment [10]. Additionally, an RT, which is purely deterministic and only depends on the geometry of the scenario,

This work was partially supported by NIST under Awards No. 70NANB18H273 and 60NANB20D082. Mattia Lecci's activities were supported by *Fondazione CaRiPaRo* under the grants "Dottorati di Ricerca" 2018.

can be combined with stochastic models for the generation of diffuse components to create a *Quasi-Deterministic (QD)* model. These depend on the roughness of the surface on which rays reflect, and are clustered around the main reflected component [15]. The modeling of these components is relevant at mmWaves as the wavelength approaches the scale of the surface roughness [9].

QD models for mmWaves have been introduced in [10], [15]. These papers, however, discuss the measurement process and the derivation of the parameters for the model, but only give a high-level overview of the mathematical formulation of the QD model. The goal of this paper is to fill that void, namely to provide the mmWave research community with a detailed recipe on how to generate realizations of NIST's implementation of the IEEE 802.11ay QD model. We will discuss the generation of a channel instance step by step, precisely describing the parameters and random distributions, using an open source QD implementation developed by NIST and the University of Padova as a reference[1]. Additionally, we will compare channels generated using this QD model with real measurements in an indoor environment at 60 GHz, to validate the accuracy of the model.

The rest of this paper is organized as follows. Section II introduces the notation that will be used throughout the paper. Section III reports the mathematical model, with the comparison in Section IV. Finally, Section V concludes the paper.

II. Notation

In the remainder of this paper, simple math font (e.g., a) is used for both scalar and vector variables, while bold math font is used for random variables (e.g., $\boldsymbol{a}$). The function $d(x_1, x_2)$ corresponds to the euclidean distance between points x_1 and x_2 in 3D space. The following notation and distributions for random variables are assumed:

- $\mathbf{X} \sim \mathcal{N}(\mu, \sigma^2)$: Normal distribution with $\mathbb{E}[\mathbf{X}] = \mu$ and $\mathrm{var}(\mathbf{X}) = \sigma^2$
- $\mathbf{X} \sim \mathcal{R}(s, \sigma)$: Rician distribution where $s, \sigma \geq 0$. It can be generated as $\mathbf{X} = \sqrt{\mathbf{Y}^2 + \mathbf{Z}^2}$, where $\mathbf{Y} \sim \mathcal{N}(s, \sigma^2)$, $\mathbf{Z} \sim \mathcal{N}(0, \sigma^2)$.
- $\mathbf{X} \sim \mathcal{L}(\mu, \sigma^2)$: Laplacian distribution with $\mathbb{E}[\mathbf{X}] = \mu$ and $\mathrm{var}(\mathbf{X}) = \sigma^2$
- $\mathbf{X} \sim \mathcal{E}(\lambda)$: Exponential distribution with $\mathbb{E}[\mathbf{X}] = \frac{1}{\lambda}$ and $\mathrm{var}(\mathbf{X}) = \frac{1}{\lambda^2}$
- $\mathbf{X} \sim \mathcal{U}[a, b]$: Uniform distribution in the closed interval $[a, b]$

III. Mathematical Model

In this section, we will provide a step-by-step tutorial on how to generate a channel with a QD model, with a precise and rigorous mathematical formulation.

The QD model considers as a basis a deterministic channel, which can be computed through ray tracing for time t, given an environment geometry, and Transmitter (TX) and Receiver (RX) positions [14]. The computed Deterministic Rays (D-rays) will then be the baseline for the multipath components randomly generated by the QD model. If present, the direct ray is treated separately as it does not generate any diffuse component.

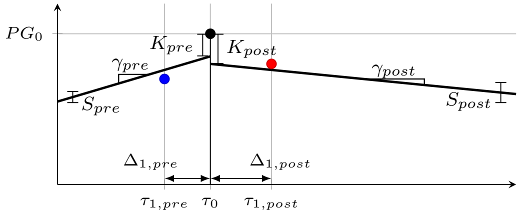


Figure 1: Graphical representation of QD parameters.

The QD model can be realized from the model for a first-order reflection and from it generalized to higher-order reflections. For reasons that will become clear later on, we define the instant in which the direct ray should arrive at the RX (even if it is actually blocked) as $t_0 = t + t_{\mathrm{dir}}$, where $t_{\mathrm{dir}} = \frac{d(\mathrm{TX},\mathrm{RX})}{c}$, and c is the speed of light. From now on we will consider a frame of reference in the variable τ relative to time t_0, where $\tau = 0$ corresponds to t_0. Given this choice, the direct ray, if it exists, will arrive at time $\tau = 0$, whereas the reflected D-rays will arrive at times $\tau > 0$.

A. First-order reflections

In this section, we will provide a step-by-step tutorial on how to generate a channel with a QD model, with a precise and rigorous mathematical formulation. The pseudo-code for this algorithm is reported in Algorithm 1, while a graphical representation of the parameters is shown in Fig. 1.

Statistics for all rays are assumed independent of their arrival time. We thus consider, without loss of generality, a single reflected D-ray with arrival time $\tau_0 > 0$, path gain PG_0, Angle of Departure (AoD) along the azimuth/elevation axes $AoD_{az/el,0}$, and Angle of Arrival (AoA) $AoA_{az/el,0}$. The same procedure will be repeated for all other reflected D-rays.

A cluster can be defined as the set with a D-ray and the corresponding MPCs. The total number of MPCs of a given cluster will be $N_{MPC} = N_{pre} + 1 + N_{post}$, including pre-cursors (i.e., diffuse components that are received before the D-ray), main cursor (i.e., the D-ray), and post-cursors (i.e., received after the D-ray). Based on some experimental evidence, we suggest to use $N_{pre} = 3$ and $N_{post} = 16$, although these numbers may vary in different locations and for different models.

The arrival times of the MPCs are modeled as a Poisson process, meaning that their inter-arrival times are independent and exponentially distributed. Namely, the post-cursors arrival times $\boldsymbol{\tau}_{i,post}$ are random variables generated based on inter-arrival delays $\boldsymbol{\Delta}_{i,post} = \boldsymbol{\tau}_{i,post} - \boldsymbol{\tau}_{i-1,post}$ as follows

$$\boldsymbol{\Delta}_{i,post} | \boldsymbol{\tau}_{i-1} \sim \mathcal{E}(\boldsymbol{\lambda}_{post}), \tag{1}$$

for $i = 1, \ldots, N_{post}$, where the arrival rate $\boldsymbol{\lambda}_{post} \sim \mathcal{R}(s_{\lambda_{post}}, \sigma_{\lambda_{post}})$ is a random variable itself. With slight abuse

[1] Available at https://github.com/signetlabdei/qd-realization/tree/feature/treetraversal.

Algorithm 1 Single Reflection QD Generator

```
function GetMpcsFirstReflection(Cursor: $\tau_0$, $PG_{0,D}$, $AoD_{az/el,0}$, $AoA_{az/el,0}$, Material)
    $RL \leftarrow \mathcal{R}\big(s_{RL,Material}, \sigma_{RL,Material}\big)$
    $PG_0 = PG_{0,D} - (RL - \mu_{RL})$
    PreCursors $\leftarrow$ ComputePre/PostCursors($\tau_0$, $PG_0$, $AoD/AoA_{az/el,0}$, Material)
    PostCursors $\leftarrow$ ComputePre/PostCursors($\tau_0$, $PG_0$, $AoD/AoA_{az/el,0}$, Material)
        return PreCursors, Cursor, PostCursors

function ComputePre/PostCursors($\tau_0$, $PG_0$, $AoD/AoA_{az/el,0}$, Material)
    $\lambda \leftarrow \mathcal{R}\big(s_{\lambda,Material}, \sigma_{\lambda,Material}\big)$
    $\Delta_i \leftarrow \mathcal{E}(\lambda), \quad i = 1, \ldots, N_{pre/post}$
    $\tau_i = \tau_0 \pm \sum_{j=1}^{i} \Delta_i$        ▷ Add for post-cursors, subtract for pre-cursors
    Remove pre-cursors with $\tau_i < 0$, update $N_{pre/post}$
    $K_{dB} \leftarrow \mathcal{R}\big(s_{K,Material}, \sigma_{K,Material}\big)$
    $\gamma \leftarrow \mathcal{R}\big(s_{\gamma,Material}, \sigma_{\gamma,materia}\big)$
    $\sigma_{s,Material} \leftarrow \mathcal{R}\Big(s_{\sigma_{s,Material}}, \sigma_{\sigma_{s,Material}}\Big)$
    $S_i \leftarrow \mathcal{N}\Big(0, \sigma^2_{s,Material}\Big)$
    $PG_i = PG_{0,dB} - K_{dB} - 10\log_{10}(e)\,\frac{|\tau_i - \tau_0|}{\gamma} + 10\log_{10}(e)\,S_i$
    Remove MPCs with $PG_i \geq PG_0$, update $N_{pre/post}$
    $\sigma_\alpha \leftarrow \mathcal{R}(\mu_{\sigma_\alpha}, \sigma_{\sigma_\alpha})$
    $\alpha_{AoD/AoA,az/el,i} \leftarrow \mathcal{L}\big(0, \sigma^2_\alpha\big)$
    $AoD/AoA_{az/el,i} \leftarrow AoD/AoA_{az/el,0} + \alpha_{AoD/AoA,az/el,i}$
    Wrap angles in $az = [0, 360)$, $el = [0, 180]$
    $\phi_i \leftarrow \mathcal{U}[0, 2\pi)$
        return ($\tau_i$, $PG_i$, $AoD/AoA_{az/el,i}$, $\phi_i$)
```

of notation, we consider $\boldsymbol{\tau}_{0,post} = \tau_0$, i.e., the time of arrival of the D-ray. Post-cursors arrival times are then computed as

$$\boldsymbol{\tau}_{i,post} = \boldsymbol{\tau}_{i-1,post} + \boldsymbol{\Delta}_{i,post} = \tau_0 + \sum_{j=1}^{i} \boldsymbol{\Delta}_{j,post}, \tag{2}$$

for $i = 1, \ldots, N_{post}$. Please note that random parameters such as λ_{post} should be extracted independently for each D-ray.

Pre-cursors will be similarly generated, with the difference that Eq. (2) will subtract inter-arrival delay, thus making $\boldsymbol{\tau}_{i,pre} < \tau_0$ for $i = 1, \ldots, N_{pre}$.

Since the number of pre/post-cursors was empirically extrapolated from measured data from [10], during the QD model generation some of them may not follow some basic assumptions. For example, when a D-ray has a delay τ_0 close to 0, some of its generated pre-cursors might arrive before the direct ray itself. Since this situation cannot happen in the physical reality, rays with $\tau_{i,pre} < 0$ are removed and N_{pre} is consequently updated.

The path gain of the D-ray is

$$\mathbf{PG}_0 = 20\log_{10}\left(\frac{\lambda_c}{4\pi\ell_{ray}}\right) - \mathbf{RL}_{dB}, \tag{3}$$

where λ_c is the wavelength of the carrier frequency, ℓ_{ray} is the total ray length, and $\mathbf{RL} \sim \mathcal{R}(s_{RL}, \sigma_{RL})$ is the random reflection loss factor given by the reflecting surface's material. If only the deterministic part of the ray-tracer is considered, the path gain $PG_{0,D}$ only includes the mean reflection loss μ_{RL}.

Once the arrival times $\boldsymbol{\tau}_i$ are known, the path gains for the MPCs can be computed as

$$\begin{aligned}\mathbf{PG}_{pre/post,i,\mathrm{dB}} = \boldsymbol{PG}_{0,\mathrm{dB}} - \boldsymbol{K}_{pre/post,\mathrm{dB}} + \\ - \frac{|\boldsymbol{\tau}_{i,pre/post} - \tau_0|}{\boldsymbol{\gamma}_{pre/post}}(10\log_{10} e) + \\ (10\log_{10} e)\mathbf{S}_{pre/post},\end{aligned} \tag{4}$$

where

- $\mathbf{K}_{pre/post,\mathrm{dB}} \sim \mathcal{R}\big(s_{K_{pre/post}}, \sigma_{K_{pre/post}}\big)$ is a loss factor,
- $\boldsymbol{\gamma}_{pre/post} \sim \mathcal{R}\big(s_{\gamma_{pre/post}}, \sigma_{\gamma_{pre/post}}\big)$ is the power-delay decay constant,
- $\mathbf{S}_{pre/post} \sim \mathcal{N}\Big(0, \boldsymbol{\sigma}^2_{s,pre/post}\Big)$ is the power-delay decay standard deviation, where $\boldsymbol{\sigma}_{s,pre/post} \sim \mathcal{R}\big(s_{\sigma_{s,pre/post}}, \sigma_{\sigma_{s,pre/post}}\big)$.

While $\mathbf{K}_{pre/post,\mathrm{dB}}$, $\boldsymbol{\gamma}_{pre/post}$, and $\boldsymbol{\sigma}_{s,pre/post}$ are independent across clusters, and $\mathbf{S}_{pre/post}$ is independently extracted for each MPC.

Since the main cursor is the one with the maximum PG when extracting the statistics from the measurements, MPCs with $\mathbf{PG}_{pre/post,i} \geq PG_{0,D}$ are removed, updating, in this case, $N_{pre/post}$.

Finally, the angle of departure in azimuth (and similarly the AoD in elevation and the AoAs in azimuth and elevation) of the MPCs are computed as

$$\mathbf{AoD}_{az,i} = AoD_{az,0} + \boldsymbol{\alpha}_{AoD,az,i}, \tag{5}$$

where $\boldsymbol{\alpha}_{AoD,az,i} \sim \mathcal{L}\Big(0, \boldsymbol{\sigma}^2_{\alpha_{AoD,az}}\Big)$ is the angle spread. The variance $\boldsymbol{\sigma}^2_{\alpha_{AoD,az}} \sim \mathcal{R}\Big(s_{\sigma^2_{\alpha_{AoD,az}}}, \sigma_{\sigma^2_{\alpha_{AoD,az}}}\Big)$ is itself a random variable independently extracted for each cluster.

Finally, the phase shift $\boldsymbol{\phi}_i$ due to both diffusion and Doppler shift is considered $\mathcal{U}[0, 2\pi)$ independently for each diffuse MPC.

B. Higher-order reflections

For the n^{th} reflection order, with $n > 1$, multiple heuristics can be thought of to compute the diffuse components. Unfortunately, the measurements taken and the models adopted to process them do not allow for a reliable confirmation of the proposed heuristics, but an extension to higher reflection orders is nevertheless needed for inclusion in a generic ray-tracer.

The path gain for specular rays with n reflections is extended as follows:

$$\mathbf{PG}_0 = 20\log_{10}\left(\frac{\lambda_c}{4\pi\ell_{ray}}\right) - \sum_{i=1}^{n} \mathbf{RL}_{i,dB}, \tag{6}$$

where $\boldsymbol{RL}_{i,dB} \sim \mathcal{R}(s_{RL,i}, \sigma_{RL,i})$, and $(s_{RL,i}, \sigma_{RL,i})$ refers to the statistics associated to the material of the i-th reflector of the given ray.

We propose two simple heuristics: a complete multiple reflection QD model and a reduced multiple reflection QD model.

Algorithm 2 Reduced Multiple Reflection QD Generator

```
function GetMpcsMultipleReflection(Cursor, MaterialList, MaterialLibrary)
    CursorOutput ← Cursor
    for Material $\in$ MaterialList do
        OtherMaterialsList ← MaterialList $\setminus$ {Material}
        PreCursors, PostCursors ← $\varnothing$
        CurrentPreCursors, CursorOutput, CurrentPostCursors ← GetMpcsFirstReflection(CursorOutput, Material)
        PreCursors ← Concatenate(PreCursors, OtherMaterialsReflLoss(CurrentPreCursors, OtherMaterialsList, MaterialLibrary))
        PostCursors ← Concatenate(PostCursors, OtherMaterialsReflLoss(CurrentPostCursors, OtherMaterialsList, MaterialLibrary))
      return PreCursors, CursorOutput, PostCursors

function OtherMaterialsReflLoss(Cursors, OtherMaterialsList, MaterialLibrary)
    for Cursor $\in$ Cursors do
        for Material $\in$ OtherMaterialsList do
            $RL \leftarrow \mathcal{R}(s_{RL,Material}, \sigma_{RL,Material})$
            Cursor.$PG$ ← Cursor.$PG$ + $(RL - \mu_{RL,Material})$
      return Cursors
```

Complete multiple reflection QD model: Upon the first scattering event, all components produced – both specular and diffuse – behave as independent components and their remaining paths are traced accordingly. We assume that every diffuse ray closely follows the path of the main cursor and further generates $N_{pre} + N_{post}$ diffuse MPCs at each bounce. The total number of MPCs generated by a single deterministic rays at the n-th reflection will thus be $N_{MPC} \sim (N_{pre} + 1 + N_{post})^n$.

Reduced multiple reflection QD model: In order to reduce the exponential complexity of the complete model, the reduced model neglects diffuse rays beyond the first order given their multiplicatively high attenuation. Instead, only diffuse rays generated directly by the deterministic ray are taken into account, each generated with the QD parameters associated to the impinging reflecting surface. Moreover, we assume that every diffuse component closely follows the main cursor, thus reflecting on the same reflectors (see Algorithm 2). Consequently, every reflector produces $N_{pre} + N_{post}$ diffuse components, thus yielding a maximum of $N_{MPC} \sim n(N_{pre} + N_{post}) + 1$, including the deterministic ray and possible rays discarded during their generation (see Section III-A).

IV. Comparison with Measurements

Given the structure of this QD model, every material must have a set of parameters for it to be appropriately simulated. It follows that given the CAD file of an environment, every surface must be associated with a material with all the necessary simulation parameters taken, for example, from a material library.

We report in the following tables examples of material libraries from NIST's Lecture Room, reformulating the mean and variance provided per material [10] into the s and σ parameters needed to generate the random parameters of the model. Measured data were taken from different TX positions pointing towards the center of the room, where a mobile RX sounder moved around the tables. Specifically, as shown in Fig. 2, considering the bottom-left corner as the origin $(x_0, y_0, z_0) = (0, 0, 0)$, TX$_1$ is positioned in $(2, 3, 2.5)$ m, TX$_2$ in $(8, 3, 2.5)$ m, TX$_3$ in $(8, 17, 2.5)$ m, TX$_4$ in $(2, 17, 2.5)$ m, and the RX performs a loop around the table.

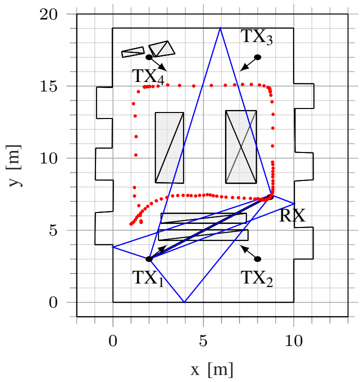


Figure 2: CAD model of NIST's lecture room. The 108 RX positions from the measurement traces are shown in red. As an example, the direct and first reflection rays generated with the RT for TX$_1$ and the specific RX position are shown in black and blue, respectively.

Given that the channel sounder's TX had a limited angular Field-of-View (FoV), it was possible to characterize different surfaces, e.g., different walls by varying the TX positions during the measurement campaign. The model parameters per position have been reformatted accordingly in Table I. Please note that, given the geometry of the room and the limited FoV, it was not possible to properly characterize some materials, in particular the floor [10]. Since no characterization was available from the measurements, no diffuse components were generated and the statistics for the reflection loss were copied from the ceiling, as this is the most similar material in the available library.

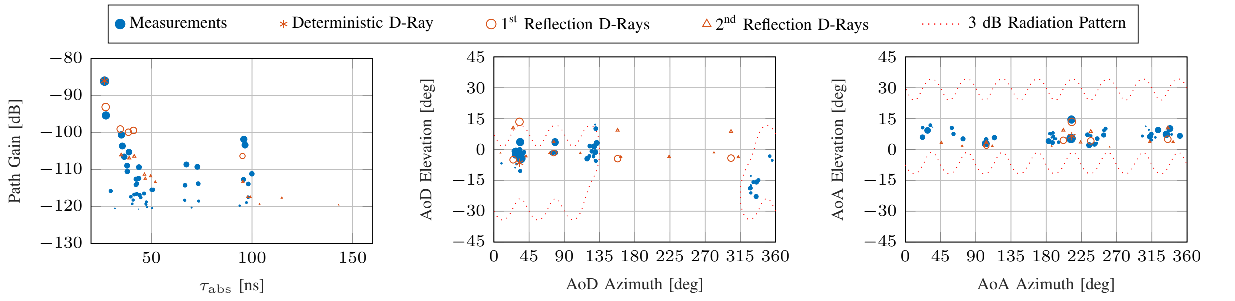
Fig. 3 shows an example of measured channel compared to the deterministic ray-traced channel for the scenario of Fig. 2. As can be seen, the direct ray is correctly identified both in the power-delay domain and in the angular domains, while other rays only partially resemble the measurements. This is due to (i) the approximate CAD model which may be missing some relevant reflectors and (ii) inaccuracies in the measurements.

While delays shown in Fig. 3a are in good accord between measurements and RT simulation, path gains are less precise, due to the random reflection losses experienced by the rays. Notice also that the TX only has antennas towards the front (as shown by the antenna pattern in Fig. 3b), thus, rays predicted by the RT to depart with an azimuth angle between 135° and 315° were not part of the real measurements. Most of all, though, it is easily noticeable that there exist clusters of rays well defined in the joint path gain, delay, AoD, AoA domain, and are missing, instead, in the channel generated by the RT. Such clusters do not arise from higher order reflections (not shown here), but rather from diffuse MPCs, thus highlighting the need for a valid diffuse QD model.

Lai, Chiehping; Wang, Jian; Gentile, Camillo; Golmie, Nada T. "Quasi-Deterministic Channel Model for mmWave: Mathematical Formalization and Validation." Presented at Workshop on ns-3 (WNS3 2021), Taipei, TW. December 07, 2020 - December 11, 2020.

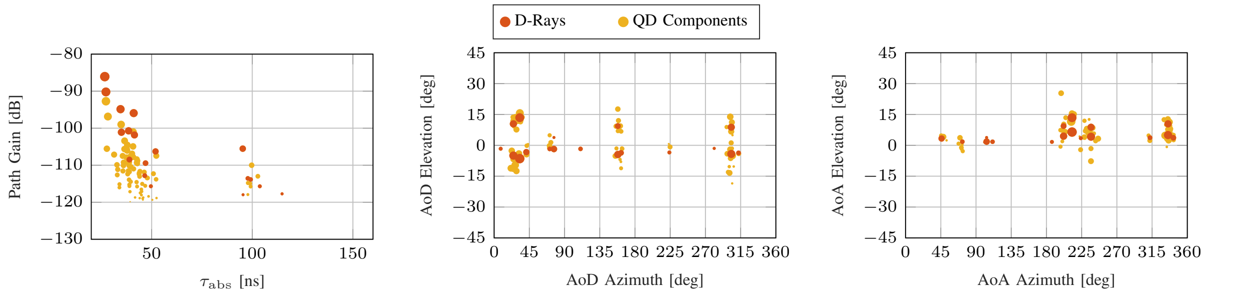
Table I: NIST's Lecture Room material library.

		Left Wall (TX$_2$)	Bottom Wall (TX$_3$)	Right Wall (TX$_1$)	Top Wall (TX$_1$)	Tables (TX$_1$)	Ceiling (TX$_1$)
$K_{dB} \sim \mathcal{R}(s, \sigma)$	$(s_{K_{pre}}, \sigma_{K_{pre}})$ $(s_{K_{post}}, \sigma_{K_{post}})$	(5.1196, 1.7485) (6.2208, 3.5421)	(1.4809, 2.1325) (7.1809, 2.5325)	(0, 0) (0.2641, 3.1699)	(0.5913, 4.5206) (0.33, 3.7213)	(0, 0) (3.7738, 1.8748)	(3.6167, 7.2715) (7.1103, 2.2712)
$\gamma \sim \mathcal{R}(s, \sigma)$	$(s_{\gamma_{pre}}, \sigma_{\gamma_{pre}})$ $(s_{\gamma_{post}}, \sigma_{\gamma_{post}})$	(0.6742, 0.9992) (0.0658, 1.2034)	(0.9006, 0.2325) (0.6881, 0.3566)	(0, 0) (0.0412, 0.8648)	(0.0094, 0.2285) (0.0792, 1.1572)	(0, 0) (0.53, 0.4837)	(0.9595, 0.901) (0.0717, 1.2794)
$\sigma_s \sim \mathcal{R}(s, \sigma)$	$(s_{\sigma_{s,pre}}, \sigma_{\sigma_{s,pre}})$ $(s_{\sigma_{s,post}}, \sigma_{\sigma_{s,post}})$	(0.0119, 0.3087) (0.4144, 0.1507)	(0.5553, 0.129) (0.26, 0.1003)	(0, 0) (0.6367, 0.3209)	(0.243, 0.273) (0.201, 0.1901)	(0, 0) (0.3309, 0.4614)	(0.2122, 0.0935) (0.7679, 0.2484)
$\lambda \sim \mathcal{R}(s, \sigma)$	$(s_{\lambda_{pre}}, \sigma_{\lambda_{pre}})$ $(s_{\lambda_{post}}, \sigma_{\lambda_{post}})$	(0.9775, 0.3449) (0.8153, 0.6948)	(0.9172, 0.2241) (1.4106, 0.5832)	(0, 0) (0.9879, 0.4235)	(0.619, 1.1299) (0.8655, 0.3762)	(0, 0) (0.8099, 0.076)	(0.8119, 0.2421) (0.7785, 0.1426)
$\sigma_\alpha \sim \mathcal{R}(s, \sigma)$	$(s_{\sigma_{\alpha,az}}, \sigma_{\sigma_{\alpha,az}})$ $(s_{\sigma_{\alpha,el}}, \sigma_{\sigma_{\alpha,el}})$	(0.1016, 2.2504) (2.9947, 1.6613)	(1.9426, 1.5726) (2.6946, 1.3948)	(3.2889, 1.3202) (3.2812, 1.8865)	(2.117, 2.1206) (2.741, 1.7964)	(1.6594, 3.1974) (4.0345, 2.6859)	(1.9829, 0.9094) (2.696, 1.1135)
$RL \sim \mathcal{R}(s, \sigma)$	(s_{RL}, σ_{RL}) μ_{RL}	(9.8412, 3.4424) 10.7	(8.5025, 4.2343) 9.84	(10.1562, 3.5164) 10.8	(6.7238, 5.9352) 9.27	(5.2106, 3.4013) 6.58	(6.5833, 2.1943) 6.9



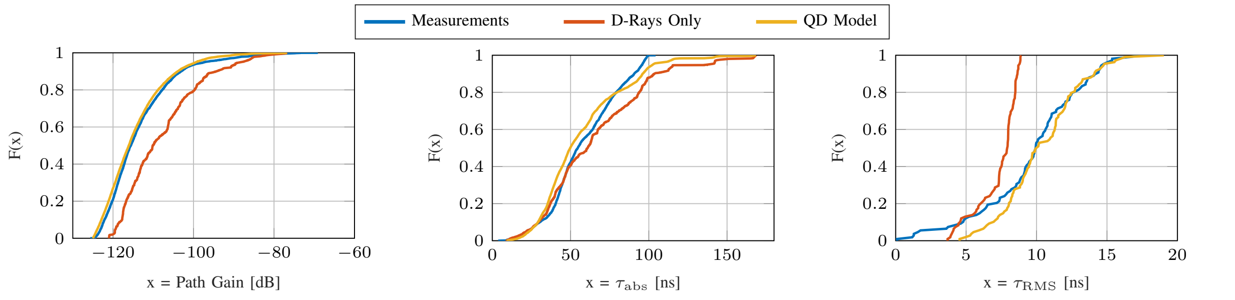
(a) Path gain vs. absolute delay (b) AoD (c) AoA

Figure 3: Example of comparison between measurements and ray-tracer, based on the channel between TX$_1$ and the RX shown in Fig. 2 in the bottom left corner of the loop. In (a), τ_{abs} represents the absolute delay of each ray. (b) and (c) show the 3 dB radiation patterns of the channel sounders described in [10] approximated with Gaussian beams. In fact, MPCs outside of these regions are not detected in the measurements.



(a) Path gain vs. absolute delay (b) AoD (c) AoA

Figure 4: Reduced multiple reflection QD model applied to RT-based channel traces with up to 2nd order reflections. Rays with path gain below -120 dB are not shown, to more closely resemble the dynamic range of the channel sounder.



(a) Path gain (b) Absolute Delay (c) RMS delay spread

Figure 5: Comparison between CDFs of MPC path gain, absolute delay, and RMS delay spread with and without QD model with respect to the measurements.

Figs. 4 and 5 show how the proposed QD model enhances the realism of a purely deterministic channel, making it significantly more similar to the measured one. Specifically, Fig. 4 reports an example of a specific channel instance, based on the CAD model shown in Fig. 2 and for the same TX/RX locations of Fig. 3. With respect to the RT specular reflections from Fig. 3, the deterministic rays (in orange), which are generated up to second order reflections, also include a random reflection loss component in the path gain. The diffuse rays added to the model are plotted in blue, with sizes proportional to the respective path gain. By comparing Fig. 4 with Fig. 3, it is clear that the D-rays alone are not able to fully model the complexity of a real channel, and that the proposed QD model can instead play an important role in this regard. In fact, empirically, rays are parts of clusters with small variations in the angular and delay domains, and large variations in the power gain domain.

Furthermore, the effects of the added rays are clearly shown in Fig. 5, which plots the Cumulative Distribution Functions (CDFs) of the path gain (Fig. 5a), the absolute delay (Fig. 5b), and the RMS delay spread (Fig. 5c), similar to the RMS angle spread shown in [10], for the multipath components of the scenarios. The CDFs show the combined statistics of the mmWave channel between TX_1 and 108 RX positions shown in red in Fig. 2). Notably, it is clear how the delays and path gains generated with the proposed QD model are significantly closer to the real measurements with respect to purely deterministic rays alone, with CDF fit improvements from 73 % to 86 % (i.e., Kolmogorov-Smirnov (KS) test improvements of 0.13) for the path gain, from 86 % to 89 % (i.e., KS test improvements of 0.03) for the absolute delay, and from 33 % to 87 % (i.e., KS test improvements of 0.54) for the RMS delay spread.

Finally, the difference in RMS delay spread, especially when the QD model is not used, can be due to the numerous reflections from objects which are not present in the CAD model but are instead part of the measured scenario, e.g., chairs and other small details of the room.

V. Conclusions

Performance evaluation is a fundamental part of the design of 5G mmWave networks. To that end, an accurate channel model allows researchers to generate reliable simulation results, that can qualitatively and quantitatively describe what can be expected when using real devices. In this paper, we introduce a mathematical formulation for a class of mmWave channels, i.e., the QD models, that can closely simulate the propagation of rays in a specific environment. We provided a step-by-step tutorial on how such models can be implemented, including the parameters and random distributions obtained from a NIST measurement campaign [10]. We then compared the results that can be obtained with an open source implementation of the model with the real measurement traces, showing improvements in the KS test for path gain (0.131), delay (0.03), and RMS delay spread (0.537).

As future work, we will further extend the QD model with material libraries from other measurement campaigns, and study methods to reduce the computational complexity involved in the ray and channel matrix generation, as in [16].

References

[1] 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, Feb. 2015.

[2] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, “End-to-End Simulation of 5G mmWave Networks,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2237–2263, Third Quarter 2018.

[3] H. Assasa, J. Widmer, J. Wang, T. Ropitault, and N. Golmie, “An Implementation Proposal for IEEE 802.11ay SU/MU-MIMO Communication in ns-3,” in *Proceedings of the 2019 Workshop on Next-Generation Wireless with ns-3*, Jun. 2019, pp. 26–29.

[4] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, “An E2E Simulator for 5G NR Networks,” *Simulation Modelling Practice and Theory*, vol. 96, Nov. 2019.

[5] 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)*, Jun. 2018.

[6] 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, Mar. 2014.

[7] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, “Millimeter wave mobile communications for 5G cellular: It will work!” *IEEE Access*, vol. 1, pp. 335–349, 2013.

[8] S. Hur, S. Baek, B. Kim, Y. Chang, A. F. Molisch, T. S. Rappaport, K. Haneda, and J. Park, “Proposal on Millimeter-Wave Channel Modeling for 5G Cellular System,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 454–469, Apr. 2016.

[9] 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.

[10] 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, Mar. 2019.

[11] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, “Millimeter-Wave Communications: Physical Channel Models, Design Considerations, Antenna Constructions, and Link-Budget,” *IEEE Communications Surveys and Tutorials*, vol. 20, no. 2, pp. 870–913, Second Quarter 2018.

[12] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, “Overview of Millimeter Wave Communications for Fifth-Generation (5G) Wireless Networks – With a Focus on Propagation Models,” *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[13] 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.

[14] 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, Dec. 2014.

[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] M. Lecci, P. Testolina, M. Giordani, M. Polese, T. Ropitault, C. Gentile, N. Varshney, A. Bodi, and M. Zorzi, “Simplified ray tracing for the millimeter wave channel: A performance evaluation,” in *Proceedings of the Workshop on Information Theory and Applications (ITA)*, Feb. 2020.

Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

A CASE STUDY OF DIGITAL TWIN FOR A MANUFACTURING PROCESS INVOLVING HUMAN INTERACTIONS

Hasan Latif

Department of Industrial & Systems Engineering
The NC State University
Raleigh, NC 27606, USA

Guodong Shao

Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA

Binil Starly

Department of Industrial & Systems Engineering
The NC State University
Raleigh, NC 27606, USA

ABSTRACT

Current algorithms, computations, and solutions that predict how humans will engage in smart manufacturing are insufficient for real-time activities. In this paper, a digital-twin implementation of a manual, manufacturing process is presented. This work (1) combines simulation with data from the physical world and (2) uses reinforcement learning to improve decision making on the shop floor. An adaptive simulation-based, digital twin is developed for a real manufacturing case. The digital twin demonstrates the improvement in predicting overall production output and solutions to existing problems.

1 INTRODUCTION

Improving shop-floor performance is critical to the future success of manufacturing. In recent years, major manufacturing countries have made strategical responses that will help make those improvements. For example, there are “Smart Manufacturing” in the USA, “Industry 4.0” in Germany, “Made in China 2025” in China, and “Industrial Value Chain Initiative” in Japan. If successful, these collective responses will enable the transition from today’s automated shop floor to tomorrow’s “smart” shop floor. “Smart” is realized by information and communication technologies and the capability to use manufacturing data for a better decision making in an integrated system of a shop floor. Such a smart shop floor will require smart production systems, smart manufacturing resources, smart manufactured products, smart raw materials, and smart human operators. Smart production systems, for example, will include order planning, production planning, job scheduling, quality control, on-time delivery, and automated fabrication. All production-system functions have two major objectives: minimize energy and minimize cost.

In a smart shop-floor, the manufacturing resources should be easily reconfigured to respond to the changing market demands and changing shop conditions (Wang et al. 2017; Helu et al. 2020). The former includes customer orders and raw materials; the latter includes the real-time status of the operators, processes, equipment, and environment. That status is used as inputs to real-time, data-analytics tools whose goal is to make optimal decisions (Zhang et. al. 2019). Despite the significant progress, generating

Latif, Hasan; Shao, Guodong; Starly, Binil. ”A Case Study of Digital Twin for a Manufacturing Process Involving Human Interactions.”
Presented at Design Automation for CPS and IoT (DESTION 2021). December 14, 2020 - December 18, 2020.

that status and making those decisions is still difficult given the typical challenges and problems on the shop-floor. The related technologies that help address these challenges include Digital Factory, Internet of Things (IoT), Cloud Computing, and Service-oriented Manufacturing System.

The aerospace, defense, and space industry are some of the top users of these technologies. Nevertheless, since their parts are normally very sophisticated and in low volume, humans are still an essential part of their production processes. There are usually about 5000 to 15 000 parts in production with different arrival times. The manufacturing process can get very complicated when it comes to a low-volume high-value product. This will add uncertainty and unpredictable breakdowns, which makes it difficult for production/factory managers to optimize or strategize the manufacturing operations. Most factory-floor inefficiencies are related to production plans. The use of digital twins will provide the production managers with timely meaningful insights to improve demand management, forecasting, schedule planning, production control, inventory management, and procurement.

There is a common misconception in the small and medium-sized enterprises (SMEs) that digitalization or digital twin modeling of their operations is too difficult or even impossible, especially in the manual process-based manufacturing industries. Therefore, the SMEs are concerned about making any significant changes in their manufacturing floor. In this paper, we demonstrate the argument made by Shao and Helu (2020) that digital twin implementations really depends on the context and viewpoint required for a specific use case, i.e., the digital twin is a fit-for-purpose digital representation by developing (1) A specific digital twin of a manufacturing process that has intensive human involvement and (2) a real case study to validate the adaptive simulation-based, digital twin.

The remainder of this paper is organized as follows: Section 2 provides some background information about digital manufacturing and digital twin. Section 3 introduces the general concept and the components of a digital twin for manufacturing processes. Section 4 discusses a case study of a manufacturing process to exemplify digital twins. Section 5 discusses the benefits and the limitation of the digital twin implementation. Section 6 concludes the paper and discuss the future work.

2 BACKGROUND

2.1 Digital Manufacturing

Digital manufacturing is an integrated approach to manufacturing that is supported by information technologies. The integration of products, processes, and resources helps manufacturers make better decisions. It not only enables data-driven, decision-support tools but also stimulates the development of new production forms such as smart manufacturing and Industry 4.0. Digital manufacturing will be essential to the creation of new solutions for manufacturing industries (Zhou et al. 2012).

Current literature in the digital manufacturing domain usually concentrate on building static models where digitalization is completely separated from the actual production floor. For increasingly complex production needs, the digitalized model lacks adaptability because of its stagnant model premise.

Digitalization provides manufacturers with more data for their products, production, and systems. Meanwhile, computing has emerged as the cheapest, most abundant resource that we can deploy to analyze that data. Stochastic simulation has been used to generate future "what if" scenarios; and, manufacturers use those scenarios to improve cost, quality, time-to-market, and throughput. In general, by combining the capabilities of data analytics, simulation, optimization, and real-time synchronization (Shao and Kibira 2018); a digital twin for manufacturing problems can be created. Of course, a specific digital twin implementation will totally depend on the scope, objective, and technologies selected for the manufacturing problem.

2.2 Digital Twin in Manufacturing

An on-going ISO standard effort defines a digital twin in the manufacturing context as "fit for purpose digital representation of an observable manufacturing element with a means to enable convergence between the element and its digital representation at an appropriate rate of synchronization (ISO 2020; Shao and Helu 2020).” The concept of a digital twin was first adopted in spacecraft design by NASA (Boschert and Rosen 2016; Brenner and Hummel 2017; Ferguson, Bennett, and Ivashchenko 2017. Grieves 2014; Grieves and Vickers 2017), viewed a digital twin as a combination of modeling-based methods and optimization-based methods. (Alam and Saddik 2017; Soderberg et al. 2017), on the other hand, as a real-time simulation with the capability of transferring information from adjacent, product, lifecycle phases.

Schleich et al. (2017) introduced a conceptual framework for building digital twins for specific applications. That framework ensured certain model properties such as scalability, interoperability, expansibility, and fidelity. Latif et al. (2019) discussed an industrial, information-integration method using Open Platform Communications Unified Architecture (OPC UA) between the real process control and a simulation of that same process. Redelinghuys et al. (2019) propose a six-layer architecture that comprises physical twin as levels 1 and 2, local data repositories as level 3, IoT gateway as level 4, cloud-based repositories as level 5, and emulation and simulation as level 6. ISO is developing a standard “Digital Twin Framework for Manufacturing” to provide a generic guideline and a reference architecture for case-specific digital twin implementations (ISO 2020).

Currently, many case studies of digital twin implementations are within a laboratory environment. For example, an Automated Guided Vehicle (AGV) or Cyber Guided Vehicle (CGV) with self-adapting behavior was developed for solving a material handling problem (Bottani et al. 2017). There are also a few industrial cases of the digital twin implementation and evaluation. For instance, Liu et al. (2018) introduce a digital twin for an automated flow-shop, Zhang et al. (2019) show a digital twin driven cyber-physical production system, and Lin et al. (2019) describe a digital twin case study for the steel industry. However, the literature about digital twin in manufacturing does not cover the manual assembly-based situation yet. SMEs are still in the rudimentary level where workers are receiving raw materials, processing parts, assembling component parts, and inspecting the final product. The digitalization of this type of process will be challenging, but once implemented, it will help eliminate non-value-added production time and other inefficiencies. The digital twin of a human-involved operation will guide the operators interactively and provide production managers with actionable information that enables them to make their decisions more effectively. This paper provides a specific prototype of a digital twin implementation for a manufacturing process with intensive human involvement, where most of the operations and related data collection are manual or semi-automatic, to demonstrate the benefits and value of a digital twin.

3 DIGITAL TWINS OF A MANUFACTURING PROCESS

This Section introduces the general concept and the components of a digital twin for manufacturing processes. Creating a digital twin for a manufacturing process starts with establishing pipelines of manufacturing data. Some of the design and manufacturing data may be collected semi or fully automatically. There are normally two kinds of manufacturing operational data: real-time and historical. The real-time operational data may be collected using smart sensors and historical data may come from existing manufacturing applications. As shown in Figure 1, a digital twin must have the following basic components: a physical element, a digital element, and their integration.

Depending on the context of the manufacturing problem and technologies selected, a digital twin may contain a variety of computational or analytic models pertaining to its real-world counterpart. Those models could range from principles-driven (natural laws), data-driven (statistical, machine learning/artificial intelligence), and geometry-driven (3D CAD), and visually-driven (virtual and augmented reality). A digital twin does not have to have all these functionalities, it should be composed entirely based on specific use case requirements (Shao and Helu 2020). In general, a digital twin can simulate the state, predict the behavior, and optimally respond to the changing conditions of its physical element through the modeling and analytics of relevant data. A feedback loop from the digital twin to the physical element, may be controlled by a user, transfers the recommendations, which provide actionable decision guidance for the production managers. This decision aid should be intuitive and help tweak or adjust the manufacturing process parameters. The process may be repeated continuously until the best-case scenario or production target is achieved.

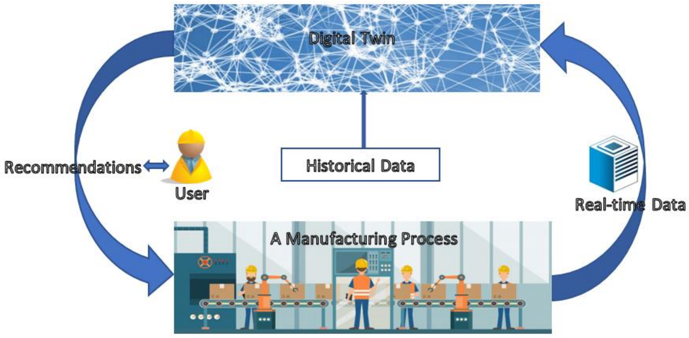


Figure 1: Concept of the Digital Twin of a Manufacturing Process

3.1 Physical Elements

In the manufacturing world, a physical element refers to the manufacturing equipment, systems, and processes on the production floor. A manual-intensive, manufacturing process may involve workers, workstations, assembly lines, and the products. The component parts arrive as raw materials. After going through all the different workstations, a final product is completed. For instance, a semiconductor process consists of raw material arrivals, fab test, die package, assembly, inspection, and delivery. Each of these elements can be the physical element of a digital twin.

3.2 Digital Elements

The digital element of a manufacturing process is the digital representation of its corresponding physical element. Depending on the use case requirements, it might include an optimization model that captures constraints and performance objectives. Or, it might include a simulation model where simulation logic and reasoning mechanisms are designed to mimic the physical operation. Or, then again, it might include a data analytics model to explore the insight of any collected data.

A database may be necessary for the digital twin to store real-time data, historical data, intermediate results, and recommendations (e.g., control commands). Data collected from the physical elements are crucial for the dynamic modeling of the manufacturing process. The data may include process parameters, product data, production-line-layout information, and information about production equipment and their operations, workpieces, material, tools, and fixtures. From a product lifecycle point of view, it may include as-designed data (product design specifications, process and engineering data), as-manufactured data (production equipment, material, method, quality, and operators), and as-maintained data (real-time and historical configuration and operation states, and maintenance records). The data may also include time and resources required to complete an operation. For a manual manufacturing processes, the data collection can also be a manual process.

3.3 Integration

A digital twin’s feasibility and effectivity entirely depend on the integration between the digital element and the physical element, i.e., the two-way communication (Shao et al. 2019). The integration should be dynamic, bi-directional, and possibly real-time.

Other than the real-time data sent from the physical element to the digital element, a feedback loop (better if automated) is needed from the digital element to the physical element for a digital twin to be successfully integrated. In reality, a feedback loop often involves human interaction, e.g., production managers select a recommendation. Once the recommendations are applied, the quality, the efficiency, the time, and the cost for production will be effectively improved.

4 THE CASE STUDY OF A DEFENSE PRODUCT ASSEMBLY

In this study, we focus on a defense product that requires a manual, assembly process and a receiving, staging area. The objective of the study is to find the best sequence of assembly operations, given the uncertainties in part-arrival times, machine-breakdown times, data-communication times, integration, and part-obsolescence times. The physical system is approximated as a linear, discrete-event, time-invariant system. The users of the digital twin are the production managers. All data are collected monthly from the production floor.

A high-level instantiation of the digital twin concept (Figure 1) is shown in Figure 2. The top half depicts the digital elements and the bottom half represents the physical elements of the manufacturing process. From a database, historical data is fed into the digital twin, which is an adaptive simulation, as initial inputs. Initial data comes into two varieties: first pass yield (FPY) and hour per unit (HPU). The digital twin model provides process recommendations as the feedback loop. Then, the production managers can select and apply the recommendations to the manufacturing process. Real-time data is fed to the digital twin from the physical manufacturing process.

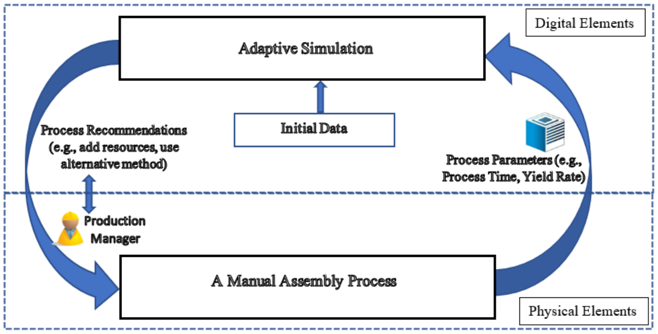

Digital Elements
Adaptive Simulation
Initial Data
Process Recommendations (e.g., add resources, use alternative method)
Process Parameters (e.g., Process Time, Yield Rate)
Production Manager
A Manual Assembly Process
Physical Elements

Figure 2: Data Flow of the Case Study

4.1 Physical Elements

The manufacturing process produces one product, called Z, which needs 44 raw materials - at different production stages - from various suppliers. Those stages involve 79 operational workstations including testing, assembling, soldering, torqueing, and inspecting. Each workstation is denoted as "opxx" where xx is the operation number. Operations along with FPY and HPU are collected for a two-year time frame starting from April 2017 to April 2019. Important assumptions about the process, for the purpose of data processing, are given below.

- Each day is 8 h and each month is 30 d.
- Only one operation can be processed at a time for an individual operation.
- There are no interruptions during the process of an individual operation, which means that the work on an operation cannot be paused in the middle and then continued later.
- All workstations are at their own location and the material is transported from one operation to another where the subsequent operation is performed. Due to missing information about transport, the time needed to transport material from one to another operation is set to zero.

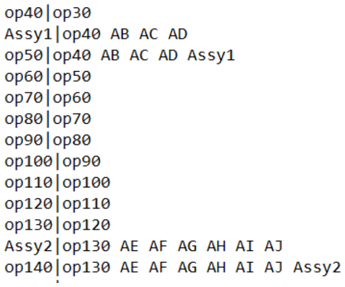
- Breaks at work, failures, troubleshooting, etc. are included in the processing times.
- Maintenance work is not considered.
- The number of workers at the workstations is not considered.
- The work orders are dependent on each other. The operations are sequential based on their operation numbers.

4.2 Digital Elements

4.2.1 Physical Process Map

Creating a digital twin of a manufacturing process requires a good understanding of the process. First, a process map needs to be created with information about all the raw materials and operations. Since the core module of the digital twin is a simulation model, the process map provides the requirements for the development and execution of the model. The key requirements include (1) the definitions of the assembly scenario to be carried out, (2) the operational data captured and analyzed for identifying the key parameters, (3) the critical process parameters and operational constraints to determine the behavior of the physical elements, and (4) the simulation of the assembly scenarios and optimization according to a set of constraints.

An “input_final” text file is prepared to capture the process map. In Figure 3, the incoming operations are provided on the right-hand side of the vertical bar and the output operation is provided on the left. The incoming raw materials are listed based on alphabetical order such as AB, AC, and AD. Operation numbers are defined by ‘opXX’ and ‘AssyX,’ where X is the operation number. For instance, op50 consists of op40; material AB, AC, AD; and assembly operation 1.



```
op40|op30
Assy1|op40 AB AC AD
op50|op40 AB AC AD Assy1
op60|op50
op70|op60
op80|op70
op90|op80
op100|op90
op110|op100
op120|op110
op130|op120
Assy2|op130 AE AF AG AH AI AJ
op140|op130 AE AF AG AH AI AJ Assy2
```

Figure 3: A section of the physical process map.

4.2.2 Data Processing

The initial dataset was taken from the first month of the data stored in an excel spreadsheet. This dataset includes operation numbers along with the time required to complete (FPY and HPU), incoming raw materials and their estimated numbers, and assembly operations. The data is pre-processed and converted into the CSV (Comma-Separated Values) format. The pre-process handles several data issues such as missing records of operations, missing operation-end timestamps, unintentional and deliberate errors, security issues related to privacy, and nondisclosure of business secrets. The clean CSV-formatted dataset is stored in a file for the simulation model to use as needed.

4.2.3 Simulation Logic

As indicated in Figure 2, the core of the digital twin is an adaptive simulation model. Traditionally, the off-line simulation has a minimal feedback loop, runs for the entire time period at once, and rarely provides assistance to the user for the next cycle. The adaptive simulation-based digital twin provides the decision-making assistance, allows real-time data input, and enables adjustment capability. The simulation runs for a specific time period (i.e., 1 month) with an initial/historical dataset (FPY and HPU for each operation). When a real-time dataset is collected, the simulation replaces that month's data with the real-time data. However, if it does not find the real-time dataset, the simulation continues with the existing dataset. The real-time dataset comes from the manufacturing process, and the data is collected, cleaned, processed, and stored in the CSV format. The simulation ends once the entire time period completes (i.e., 12 months). A recommendation list is generated by the digital twin; the list gets re-ranked constantly based on the score of each item on the list.

Figure 4 shows the simulation logic in a nutshell. When the adaptive simulation moves into the "wait" mode, it asks the users if they want to review the process for the worst performing operations. If the user wants to review the recommendation list, based on the latest information, the digital twin shows the five worst performing operations and the user selects one of them to review. Once the user selects and applies the recommendations from the list, the digital twin generates a dataset with improved parameters for that operation. Since the operations are picked from future time periods, the generated dataset waits for the actual operation to happen. Once the actual operation happens and the real-time data for that operation updates the dataset, the generated dataset is compared with the updated dataset. If the difference between the datasets stays below a pre-defined threshold value (30 % of the real-time dataset), it signifies that the recommendation has improved the process output. On the contrary, if the difference stays above the threshold value, it shows the selected recommendation has not worked better. Either way, the feedback goes to the recommendation list and re-ranks the list for the next cycle of usage.

4.2.4 Simulation Execution

The adaptive simulation model was developed using Python 3.6. When a specific operation (e.g., op560) is executed, assuming it is from the worst 5 operations, the user can compare the real-time dataset (FPY and HPU) with the generated dataset (10 % improved FPY and 10 % decreased time). After applying the recommendations, the real-time data should demonstrate significant improvement from the generated data. If the parameters (FPY and HPU) of the real-time dataset are more than 30 % of the parameters of the generated dataset, the applied recommendations will be regarded as a failure, will get a score of 0, and will get pushed down on the recommendation list. If the parameters are similar (real-time data is within 30 % of generated data), it indicates that the recommendations work well, it will get a score of 1, and will get push up in the recommendation list. In either cases, the recommendation list will be updated and stored in the database for use in the next cycle as the initial dataset. With this rule, the recommendation list for the operations always gets updated; this reflects the reinforcement learning (RL).

RL is a "trial-and-error" approach that the learning agents learn optimal decisions by interacting with the environment. The "trial-and-error" rule means RL agents make a trade-off between known decision exploitation and new decision exploration to achieve an optimal policy. Figrure 5a shows the RL model (Barto and Sutton 1997). A RL agent and its environment interact over a sequence of discrete time steps. At each time period t, the agent completes an iteration with the environment. The action a_t are the chocies made by the agent in state s_t. In the t^{th} iteration, the agent observes the current environment state s_t, and chooses an action a_t. After that, the environment transfers from the state s_t to s_{t+1} following the a state transition probability and returns a reward $r_t(s_t, a_t)$ according to the performance of a_t. So the rewards are the basis for evaluating agents' chocies. Figure 5b is an instantiaiton of the general RL concept for this case study. The environment is the recommendation list, the agent is the threshold value, the state is the simulation period, the reward is generated simulation data, and the action is to re-rank the recommendation list. In a different state, new data enters into the equation and the threshold value is compared with the generated input. A new recommendation list is created.

Eventually, all items in the recommendation list will truly represent the recommendations that have been tested and verified with a proven track record. In figure 6, the default initial recommendation list (Figure 6a) and the re-ranked recommendation list (Figure 6b) for Op560 in the next cycle are shown. The re-ranked recommendation list is showing “Find An Alternative Material” at number 1 because in the previous run, the user has selected and successfully found that the recommendation is helpful.

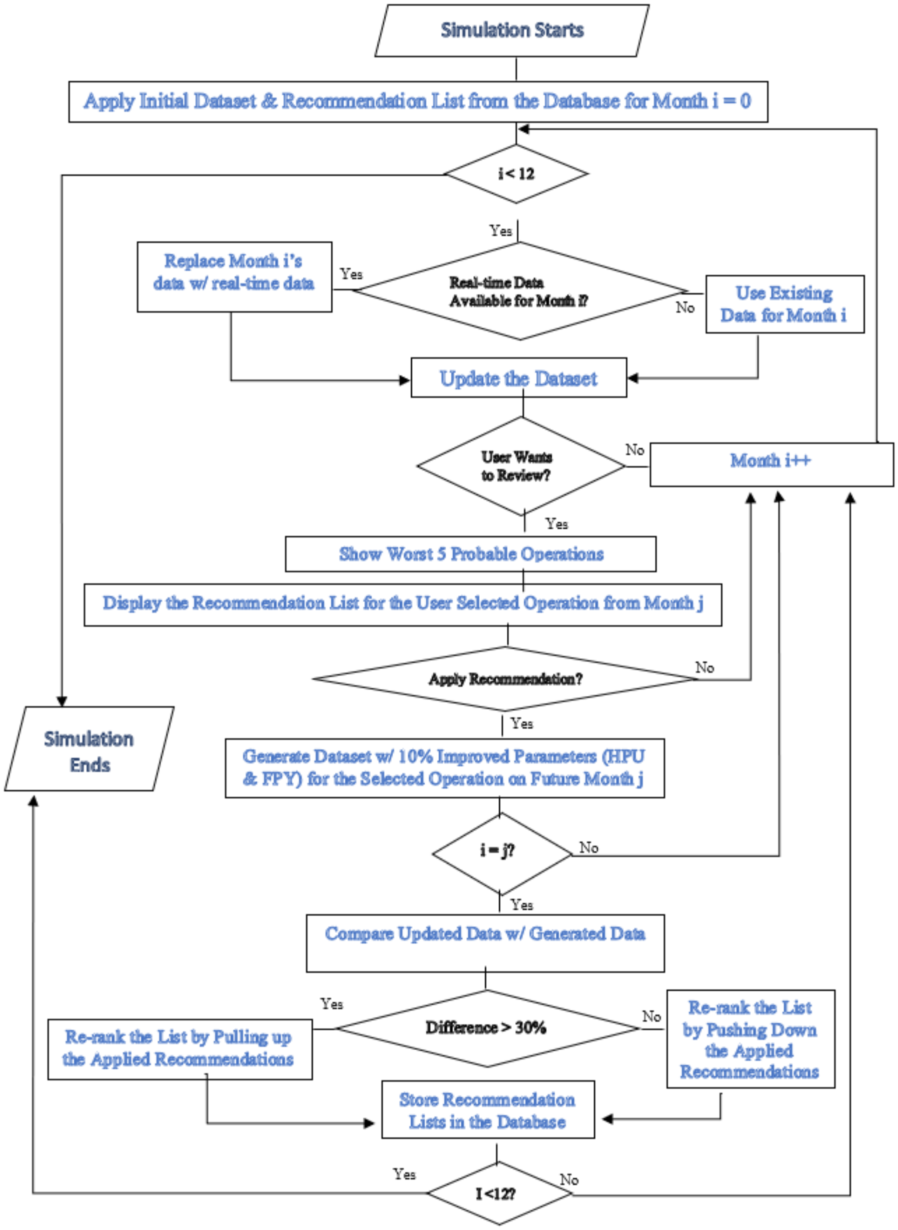

Simulation Starts
Apply Initial Dataset & Recommendation List from the Database for Month i = 0
i < 12
Yes
Replace Month i's data w/ real-time data
Yes
Real-time Data Available for Month i?
No
Use Existing Data for Month i
Update the Dataset
User Wants to Review?
No
Month i++
Yes
Show Worst 5 Probable Operations
Display the Recommendation List for the User Selected Operation from Month j
Apply Recommendation?
No
Yes
Simulation Ends
Generate Dataset w/ 10% Improved Parameters (HPU & FPY) for the Selected Operation on Future Month j
i = j?
No
Yes
Compare Updated Data w/ Generated Data
Yes
Difference > 30%
No
Re-rank the List by Pulling up the Applied Recommendations
Re-rank the List by Pushing Down the Applied Recommendations
Store Recommendation Lists in the Database
Yes
I<12?
No

Figure 4: Simulation logic for the digital twin of the manufacturing process.

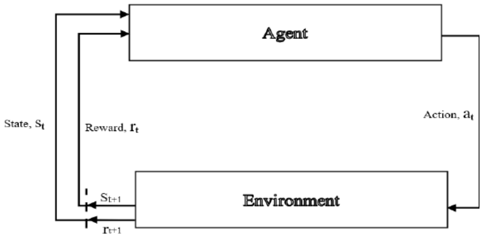


Figure 5a: The RL model.

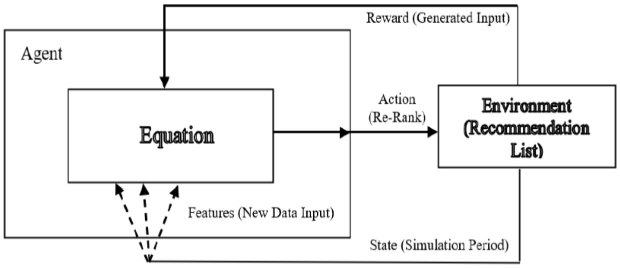
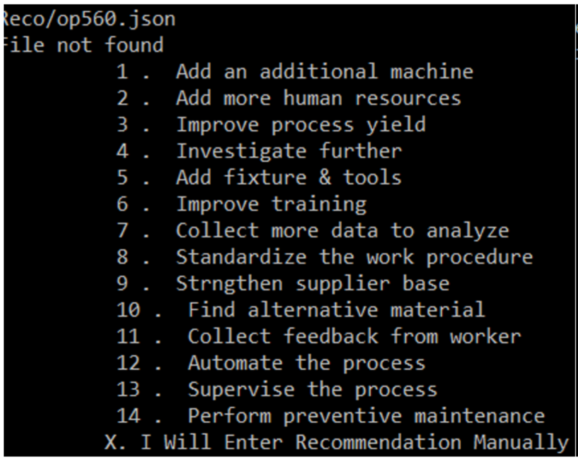
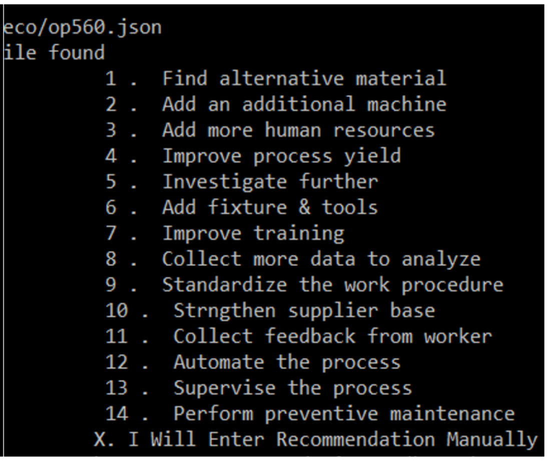


Figure 5b: RL architecture for the case study.



```
Reco/op560.json
File not found
          1 .  Add an additional machine
          2 .  Add more human resources
          3 .  Improve process yield
          4 .  Investigate further
          5 .  Add fixture & tools
          6 .  Improve training
          7 .  Collect more data to analyze
          8 .  Standardize the work procedure
          9 .  Strngthen supplier base
          10 .  Find alternative material
          11 .  Collect feedback from worker
          12 .  Automate the process
          13 .  Supervise the process
          14 .  Perform preventive maintenance
         X. I Will Enter Recommendation Manually
```

Figure 6a: Initial recommendation list.



```
eco/op560.json
ile found
        1 .  Find alternative material
        2 .  Add an additional machine
        3 .  Add more human resources
        4 .  Improve process yield
        5 .  Investigate further
        6 .  Add fixture & tools
        7 .  Improve training
        8 .  Collect more data to analyze
        9 .  Standardize the work procedure
        10 .  Strngthen supplier base
        11 .  Collect feedback from worker
        12 .  Automate the process
        13 .  Supervise the process
        14 .  Perform preventive maintenance
       X. I Will Enter Recommendation Manually
```

Figure 6b: Re-ranked recommendation list.

4.2.5 Model Validation

To validate the model, 24 months of actual data is collected. The first 12 months data is used as initial data. The rest of the 12 months data is used to validate the model. In addition to that, a desk audit has been performed to validate the output of the adaptive simulation model. In future work, formal verification and validation techniques will be applied.

4.3 Integration

The digital twin includes two-way communications: (1) the real-time data is collected from the manufacturing process, processed, and updated in the digital twin and (2) the recommendation list is generated by the digital twin and applied to the manufacturing process. In this case study, most of the data are collected manually, processed in a separate application.

A Python-based parser was developed to interpret the CSV files, a text file provides the operation sequence. The recommendation list is saved in JSON. Each time the user manually enters a recommendation or updates the recommendation list, a new JSON file replaces the older one for that specific operation. For instance, if a user makes an update on OP560, the simulation first looks for a JSON file in the recommendation list. If it does not find the JSON file, it shows the default recommendation list. Based on user's selection, the recommendation list gets updated and saved in the OP560 JSON file. Next time, whenever the user wants to see the recommendation list for OP560, the OP560 JSON file is shown for the user. The Python parser does all the data transfer except the initial CSV file generation. The data is transferred once in every month, so the overall effort to run the digital twin is minimal. Even though the data transfers monthly, the data comes from the real-manufacturing floor, and therefore, it is considered as real-time data.

In this case study, the initial dataset comes from the historical database. It is the data from April 2017 to March 2018. The simulation will ask for the real-time data for each month, e.g., April 2018, May 2018, or June 2018, once the real-time data is available, simulation replaces the existing dataset with the real-time data, and recalculate the output at the end of the simulation. If the user wants to improve the efficiency or reduce the processing time of the operation, the simulation provides the worst, five, performing operations that have the highest potential to be improved and a recommendation list for the improvement of each operation. The equation used to find the worst five operations is given below.

$$x = HPU * FPY^2 \ . \tag{1}$$

For instance, OP560 has two parameters: FPY as 1.0 and HPU as 0.75. OP560's 'x' value calculates as 0.75. With a similar calculation, the top, five, worst operations are derived based on the lowest 'x' value. Therefore, operations with a lower yield rate and higher HPU will show up as a worse performing operation.

The collected data is fed into the digital twin to replace the previous dataset. Therefore, the digital twin can periodically adjust its prediction output and point out the areas to improve. The recommendation list brings useful insight to the worst operations. Based on the application and processed data feedback, the recommendation list constantly gets evolved. The successful recommendations come at the top of the list and the failed ones go to the bottom of the list. Therefore, every operation shows a unique list that has been tested and verified over the period of time. For instance, after multiple runs of the digital twin, a recommendation list can show the recommendations that have been successful for the similar situation multiple times. Therefore, the user can make a better decision.

5 DISCUSSION

The digital-twin concept is still new in the manufacturing domain. There are not many successful digital twin implementations available especially for those processes that have a lot of human interactions (Ma et al. 2019). In addition to the confusion of the digital twin concept, the advent of new information technologies also plays a role in this, because designing, implementing, and integrating digital twins with those technologies could get very complicated. The idea of a digital twin has to be a comprehensive virtual replication of all physical and functional activities within a shop floor makes it difficult for manufacturers to adopt, invest, and implement digital twins. Completing specific use cases of digital twins with a manageable scope will help them better understand the efforts they need to involve and benefit they will get. This case study provides a success story of digital twin implementation even for a manual, complicated, shop-floor problem with uncertainties associated with multiple materials, operation sequences, human interactions, and real-world data. The case also provides an implementation procedure.

Part of that procedure involves creating a process map. Other parts include data collection, data processing, and data integration. Users (i.e., production managers) need to understand the impact of these data-related issues on the digital twin output. In addition, since most of the operations involve manual assembly, managers need to understand how process variability contributes to the variability in the data. In our case, we simply used the average of the data to nullify that variability. However, an advanced level of mathematical technique could be deployed to tackle the variability problem.

Because of all these uncertainties, a dynamic simulation model is built. The simulation output is a recommendation list that includes what to make, when to make it, and how to make it. The recommendation list evolves continuously. Although the recommendation list should reflect the appropriate solutions for the manufacturing operations, the impact of each recommendation has been assumed equal. In real-life the impact is obviously different, however, for simplicity, the difference between impacts has not been accounted. Similarly, the threshold value for the comparison of the two parameters between the generated dataset and real-time data set should be different based on the operations. For implementation feasibility, the threshold point is kept the same for all the operations.

6 CONCLUSION AND FUTURE WORK

The digital twin can provide concrete value and help production managers take the key strategic decisions. A digital twin can have many applications across its life cycle depending on its context and purpose. It can answer the critical what-if questions more accurately in real-time. It is expected that the applications of digital twins can contribute to improved operations management, operations execution, resource utilization, lead times, and due-date reliability.

In this paper, an adaptive, simulation-based, digital twin has been implemented. The real, case study showed proof of concept. It further proves that digital twin needs to be use case specific. Specifically, this work targets a human-involved manufacturing process where automation is not completely available. It can be speculated that this approach will have instructional significance when manufacturing industries build or revamp a plant, arrange the facilities or staff, and polish up the process flow.

With more data and applications, further issues and challenges of the development, implementation, validation of the digital twin in real manufacturing environments will be identified. As a future activity, the authors would like to apply the approach to more manufacturing problems including a more detailed evaluation of the machine’s health and the planning of the maintenance activities.

ACKNOWLEDGMENTS

The authors are indebted to the factory managers of Raytheon Technologies for providing insightful information to this research.

DISCLAIMER

In the case study, all data points and physical process maps are masked and do not represent the actual operations. 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. No approval or endorsement of any commercial product by NIST is intended or implied.

REFERENCES

Alam, K. M. and A. E. Saddik. 2017. “C2PS: A Digital Twin Architecture Reference Model for the Cloud-based Cyber-Physical Systems”. *IEEE Access* 5: 2050-2062.

Barto, A. G. and R. S. Sutton. 1997. “Reinforcement Learning in Artificial Intelligent”. In *Neural-Networks Models of Cognition,* edited by J. W. Donahoe and P. Dorsel, 358–386. Biobehavioral Foundations.

Boschert, S. and R. Rosen. 2016. “Digital Twin - the Simulation Aspect”. In *Mechatronic Futures: Challenges and Solutions for Mechatronic Systems and Their Designers*, edited by P. Hehenberger and D. Bradley, 59-74. Cham: Springer.

Bottani, E., Cammardella, A., Murino, T., and S. Vespoli. 2017. “From the Cyber-Physical System to the Digital Twin: the Process Development for Behaviour Modelling of a Cyber Guided Vehicle in M2M Logic”. *XXII Summer School Francesco TurcoIndustrial Systems Engineering*, Sept. 13th-15th, Palermo, Italy, 1-7.

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.

Ferguson, S., Bennett, E., and A. Ivashchenko. 2017. “Digital Twin Tackles Design Challenges”. *World Pumps* 2017(4): 26-28.

Grieves, M. 2014. “Digital Twin: Manufacturing Excellence through Virtual Factory Replication”. White paper 1: 1-7. Florida Institute of Technology.

Grieves, M., and J. Vickers. 2017. “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems”. In *Transdisciplinary Perspectives on Complex Systems*, edited by F. J. Kahlen, S. Flumerfelt, and A. Alves, 85-113. Cham: Springer.

Helu, M., Sobel, W., Nelaturi, S., Waddell, R., and S. Hibbard. 2020. “Industry Review of Distributed Production in Discrete Manufacturing.” *Journal of Manufacturing Science and Engineering* 142(11): 111001-111014.

ISO. 2020. “ISO (DIS) 23247-1: Automation Systems and Integration - Digital Twin Framework for Manufacturing - Part 1: Overview and general principles”. ISO/TC 184/*SC4*/WG15.

Latif, H., Shao, G., and B. Starly. 2019. “Integrating a Dynamic Simulator and Advanced Process Control using the OPC-UA Standard”. *Procedia Manufacturing* 34: 813-819.

Li, C., Mahadevan, S., Ling, Y., Wang, L., and S. Choze. 2017. “A Dynamic Bayesian Network Approach for Digital Twin”. In *proceedings of the 19th AIAA Non-Deterministic Approaches Conference*: 1566.

Lin, S. W., Fu, M., and K. Li. 2019. “Digital Twin + Industrial Internet for Smart Manufacturing: A Case Study in the Steel Industry”. The Industrial Internet Consortium, Needham, MA.

Liu, Q., Zhang, H., Leng, J., and X. Chen. 2019. “Digital Twin-Driven Rapid Individualised Designing of Automated Flow-Shop Manufacturing System”. *International Journal of Production Research* 57(12): 3903-3919.

Ma, X., Tao, F., Zhang, M., Wang, T., and Y. Zuo. 2019. “Digital Twin Enhanced Human-Machine Interaction in Product Lifecycle”. *Procedia CIRP* 83: 789-793.

Redelinghuys, A. J. H., Basson, A. H., and K. Kruger. 2019. “A Six-Layer Architecture for the Digital Twin: a Manufacturing Case Study Implementation”. *Journal of Intelligent Manufacturing*: 1-20.

Schleich, B., Anwer, N., Mathieu, L., and S. Wartzack. 2017. “Shaping the Digital Twin for Design and Production Engineering. *CIRP Annals* 66(1): 141-144.

Shao, G., and M. Helu. 2020. “Framework for a Digital Twin in Manufacturing: Scope and Requirements”. *Manufacturing Letters* 24: 105-107.

Shao, G., Jain, S., Laroque, C., Lee, L. H., Lendermann, P., and O. Rose. 2019. “Digital Twin for Smart Manufacturing: The Simulation Aspect”. 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, 2085-2098. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Shao, G., and D. Kibira. 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.

Söderberg, R., Wärmefjord, K., Carlson, J. S., and L. Lindkvist. 2017. “Toward a Digital Twin for Real-Time Geometry Assurance in Individualized Production”. *CIRP Annals* 66(1): 137-140.

Wang, C., Jiang, P., and K. Ding. 2017. “A Hybrid-Data-on-Tag–Enabled Decentralized Control System for Flexible Smart Workpiece Manufacturing Shop Floors”. In *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 231(4): 764-782.

Zhang, H., Zhang, G., and Q. Yan. 2019. “Digital Twin-Driven Cyber-Physical Production System towards Smart Shop-Floor”. *Journal of Ambient Intelligence and Humanized Computing* 10(11): 4439-4453.

Zhou, Z., Xie, S. S., and D Chen. 2011. *Fundamentals of Digital Manufacturing Science*. London: Springer-Verlag.

AUTHOR BIOGRAPHIES

HASAN LATIF works as a Sr. Manufacturing Engineer at Missiles & Defense Systems in Raytheon Tehnologies. He is a pasisonate researcher in the manufacturing domain. His research interest is in smart manufacturing, simulation, and industry 4.0, and he has relevant research experiences in the application areas of aerospace & defense industrty. He has a Ph.D. in Indsutrial Engineering from The NC State University. His email address is hhlatif@ncsu.edu.

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 twins; 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.

BINIL STARLY is a professor in the Industrial & Systems Engineering Department at North Carolina State University. He directs the Data Intensive Manufacturing Laboratory (DIME Lab). His specific technical expertise is in digital design and fabrication, and cyber-physical systems in manufacturing. Binil has received the National Science Foundation CAREER award for research in multi-scale biological tissue scaffold systems built from additive manufacturing platforms. His email address is bstarly@ncsu.edu.

A Visual Analytics Approach for the Diagnosis of Heterogeneous and Multidimensional Machine Maintenance Data

Xiaoyu Zhang[*] University of California, Davis
Takanori Fujiwara[*] University of California, Davis
Senthil Chandrasegaran[†] Delft University of Technology
Michael P. Brundage[‡] National Institute of Standards and Technology
Thurston Sexton[‡] National Institute of Standards and Technology
Alden Dima[‡] National Institute of Standards and Technology
Kwan-Liu Ma[*] University of California, Davis

Abstract

Analysis of large, high-dimensional, and heterogeneous datasets is challenging as no one technique is suitable for visualizing and clustering such data in order to make sense of the underlying information. For instance, heterogeneous logs detailing machine repair and maintenance in an organization often need to be analyzed to diagnose errors and identify abnormal patterns, formalize root-cause analyses, and plan preventive maintenance. Such real-world datasets are also beset by issues such as inconsistent and/or missing entries. To conduct an effective diagnosis, it is important to extract and understand patterns from the data with support from analytic algorithms (e.g., finding that certain kinds of machine complaints occur more in the summer) while involving the human-in-the-loop. To address these challenges, we adopt existing techniques for dimensionality reduction (DR) and clustering of numerical, categorical, and text data dimensions, and introduce a visual analytics approach that uses multiple coordinated views to connect DR + clustering results across each kind of the data dimension stated. To help analysts label the clusters, each clustering view is supplemented with techniques and visualizations that contrast a cluster of interest with the rest of the dataset. Our approach assists analysts to make sense of machine maintenance logs and their errors. Then the gained insights help them carry out preventive maintenance. We illustrate and evaluate our approach through use cases and expert studies respectively, and discuss generalization of the approach to other heterogeneous data.

Keywords: Visual analytics, heterogeneous data, high-dimensional data, machine learning, text analytics, maintenance logs.

1 Introduction

Making sense of large-scale, heterogeneous data is one of the main challenges faced by data science and visualization communities in real-world application scenarios. For instance, in large-scale manufacturing setups, human- and machine-created logs of operation and maintenance need to be analyzed to identify problem areas and prevent major failures before they occur [8]. These logs can easily number over hundreds of thousands of records and often include multiple types of data: numerical data (e.g., operating temperatures), categorical data (e.g., machine types), ordinal data (e.g., error severity), and text data (e.g., machine status description) [26]. In addition, such logs also feature manual entries—including natural-language descriptions—which are prone to inconsistencies, such as the same problem described differently at different times or by different people [44]. These factors make it difficult for managers and technicians—even with the help of data analysts—to analyze logs to identify patterns (e.g., common phenomena seen in some type of errors) and perform preventive maintenance. While such issues are common in maintenance analysis and prognostics, the challenge of heterogeneous and inconsistent data spans domains.

Machine learning (ML) assisted visual analytics has been developed to address the challenge in reviewing large, high-dimensional data [42, 52]. For instance, researchers have used dimensionality reduction (DR) to provide an overview of high-dimensional data in lower dimensions [28, 54] and clustering to summarize the information of large data into a small number of groups [3, 29]. Contrastive learning, which extracts salient patterns in one dataset relative to the other, is then used to help interpret the results of DR and clustering [20, 22]. Such approaches can help maintenance log analysts extract and explain important patterns specific to certain kinds of issues, while data inconsistency can be mitigated by keeping the human in the loop. However, these ML methods are designed to apply to a single datatype, such as numerical or categorical. Thus, when analyzing heterogeneous data, we need to consolidate different methods. In addition, existing contrastive learning methods are applicable only to either numerical or binary data. New methods for other datatypes (e.g., categorical and text) are needed.

In this paper, we present an approach to separate different variable types—numerical, categorical, and text—in a heterogeneous dataset and provide lower-dimensional, clustered visualizations for each type. We then use ccPCA [20]—contrasting clusters in Principal Component Analysis—as the contrastive learning method for *numerical* variables in the data. To provide a similar functionality for *categorical* variables, we introduce a method called contrasting clusters in Multiple Correspondence Analysis (ccMCA). ccMCA helps characterize a selected cluster (of categorical data) by comparing its attributes with those of the remaining data. For *text* variables, we first convert natural-language descriptions into high-dimensional vectors using word embeddings [48], and then perform DR and clustering. In place of contrastive learning, we plot text frequencies compare each cluster with the rest of the data.

Finally, we link the visualizations across all the views to help the analyst characterize clusters in the context of the other data dimensions. We illustrate our approach with use-case scenarios and expert reviews using a real-world dataset of maintenance and repair logs for heating, ventilation, and air-conditioning (HVAC) systems.

Our main contributions include: (1) integrating existing DR and clustering techniques to make sense of multidimensional, heterogeneous maintenance log data by introducing a visual analytics approach to coordinate views resulting from these techniques for each datatype, (2) introducing a new contrastive learning method called ccMCA to help the user characterize data clustered on the basis of categorical dimensions, and (3) illustrating the use of domain knowledge to characterize the clustered data.

2 Related Work

While the proposed work falls under the application area of machine maintenance data analysis, our approach draws from and contributes to existing approaches in heterogeneous and high-dimensional data. We highlight representative research on these topics in this section.

[*] e-mail: {xybzhang, tfujiwara, klma}@ucdavis.edu
[†] e-mail: r.s.k.chandrasegaran@tudelft.nl
[‡] e-mail: {michael.brundage, thurston.sexton, alden.dima}@nist.gov

Zhang, Xiaoyu; Fujiwara, Takanori; Chandrasegaran, Senthil; Brundage, Michael; Sexton, Thurston; Dima, Alden; Ma, Kwan-Liu. "A Visual Analytics Approach for the Diagnosis of Heterogeneous and Multidimensional Machine Maintenance Data." Presented at 14th IEEE Pacific Visualization Symposium (PacificVis 2021). April 19, 2021 - April 21, 2021.

2.1 Machine Maintenance Log Analysis

With an increasing emphasis on smart manufacturing and reducing machine downtime, process monitoring, diagnostics, and prognostics have gained prevalence. This trend—coupled with cheaper and more accessible sensors and data storage solutions—has led to an increase in maintenance data [8]. Despite the potential benefits of high-volume maintenance data for better machine management, companies frequently struggle to adopt advanced manufacturing technologies and strategies due to cost and lack of technical expertise in data analysis [27]. Simple yet powerful solutions for data analysis are necessary to aid manufacturers in improving their practices. There has been an increasing focus on sensor data and predictive maintenance using AI techniques [11,51]. However, these works often neglect a large portion of maintenance data: natural language, short-text maintenance logs. Annotation methods for short-text maintenance work orders [34,43] have been the subject of recent research. For instance, Sexton et al. [45] developed Nestor (https://nist.gov/services-resources/software/nestor), an open-source tool that uses internal "importance" heuristics and seed data annotated with domain-relevant tags by experts. Nestor uses these to annotate maintenance logs with similar tags.

Visual analytics is another technique that has gained popularity in this domain in recent years. Notable work in visual analytics for machine log visualization and monitoring includes ViDX [56] for historical analysis and real-time monitoring of assembly lines, La VALSE [24] and MELA [46] for interactive event analysis logs, and ViBR [12] for vehicle fault diagnostics. These approaches are created for specific datatypes. On the other hand, we treat the data as high-dimensional, heterogeneous datasets that include unstructured text, making our approach usable across different domains.

2.2 Visualizing Heterogeneous Data

The challenges of visualizing heterogeneous data, i.e., data with mixed datatypes or variables, such as numerical, categorical, and text, were recognized early in visualization research. Almost 25 years ago, Zhou and Feiner [57] provided a systematic approach to design visualizations for heterogeneous data based on data characteristics and the tasks involved. The *size* of heterogeneous datasets poses additional challenges for visualization, such as requiring large screens and appropriate visual mappings. Different approaches were developed to address these challenges, such as developing automated specification algorithms to map data attributes to visual attributes [9], and high-resolution immersive visualization environments [40].

Visualizing heterogeneous data also provides a way for the user to establish *context*. For instance, coordinated timeline visualizations of audio, video, and text data of human-human or human-machine interactions can provide context to observations about movement, speech, and activity data [13,19]. More recently, immersive visualizations of system activity overlaid on a spatial layout corresponding to the physical locations of said systems were used to provide contextual information in real-time network security analysis [35].

Unstructured text also forms an important datatype. Descriptive text about problems and repairs is often entered by operators and maintenance personnel who assume familiarity with the machines and related processes. The text thus tends to be terse and laden with jargon, and is often inconsistent across people. Developing a lexicon—a domain-specific vocabulary—is often necessary to interpret such text data in a semantically consistent way. The General Inquirer [50] is one of the earliest attempts to build a lexicon for content analysis of text. Categories such as Linguistic Inquiry and Word Count (LIWC) [39] focus on psychological relevance (such as moods) and general-purpose applications. Such models are trained on general text corpora such as news articles, online forums, and fiction. For application to large-scale technical text data, automated tagging needs to be balanced via manual sifting of the text.

Visual analytics has been used to achieve better-balanced tags, using a combination of high-dimensional data visualizations and user-steered analyses. For instance, ConceptVector [37] visualizes word-to-concept similarities to guide users to categorize text data given a specific domain, such as politics or finance. Similar vector space representations are used by Heimerl and Gleicher [25] to design visualizations that help users understand word vector embeddings. In addition, several tools such as the Exploratory Labeling Assistant [18] and AILA [14] use machine-learning based recommendations to help users characterize or label documents.

Drawing from this combination of statistical and manual approaches, we use word embeddings to translate short texts to high-dimensional vectors, and apply DR and clustering to find groups of semantically related short texts in a 2D space. We use similar DR and clustering representations for numerical and text data dimensions, which gives us consistent representations across datatypes.

2.3 Visualizing High-Dimensional Data

Most machine maintenance log data tend to be high-dimensional, as each breakdown or maintenance event is recorded with multiple fields relating to different personnel and/or departments [44]. While high dimensionality has its advantages, such as the ability to contextualize and correlate features of the data, it also makes the data less usable for sampling or statistical analysis [15]. Dimensionality reduction provides a lower-dimensional representation while preserving the essential information of the original data [54]. Nonlinear DR techniques, such as UMAP [36], are especially relevant for large-scale, high-dimensional data as they preserve local neighbor relationships, which can help identify subgroups in the data.

DR can be further exploited to cluster the data with higher speed and performance [47] or to produce an overview of the data [32,42]. During this process, visual analytics of the clustered data is often needed to help users determine *which* attributes contribute to the distinctness of each cluster [5]. Statistical charts (e.g., boxplots) [29] or density plots [49] of selected clusters from the DR result have been used for this purpose. However, showing one statistical chart for each attribute becomes visually overloaded as the number of attributes increases. A better approach would be to identify and visualize salient attributes that contribute to a selected cluster. For instance, Broeksema et al. [6] visualized the results of multiple correspondence analysis (MCA) [30]—a variant of principal component analysis (PCA) for categorical data—together with a colored Voronoi cell that represents a highly-related attribute to each data point. Similarly, Joia et al. [28] drew a convex hull around each cluster and filled the resulting polygon with a word cloud consisting of names of the attributes related to the cluster. Faust et al. [17] took a different approach, using local perturbations in the input data to represent how the higher dimensions are represented in the projected views. More recently, Fujiwara et al. [20] used contrastive learning to find attributes that contrast a selected cluster from the rest of the data. We incorporate this contrastive learning-based approach to analyze the numerical attributes of the maintenance log data while introducing an analogous approach for the categorical attributes.

3 Requirements

Typically, visual analysis of heterogeneous, multidimensional data is performed with the goal of identifying patterns within the data and extracting meaning from them [2, 55]. With our application area of machine maintenance data analysis in mind, we draw our requirements from existing work on maintaining and tagging machine performance, error, and maintenance log data.

Most of our requirements are based on prior work by Brundage et al. [7, 8] who generate a set of commonly-occurring data elements from their study of various maintenance work order datasets including temporal (e.g., time between failures, machine downtime, etc.), machine (machine type, location, etc.) human (operator/tech name, skill level, etc.), raw text (problem descriptions, solution, etc.), and tagged elements (items, actions, etc.). Broadly speaking,

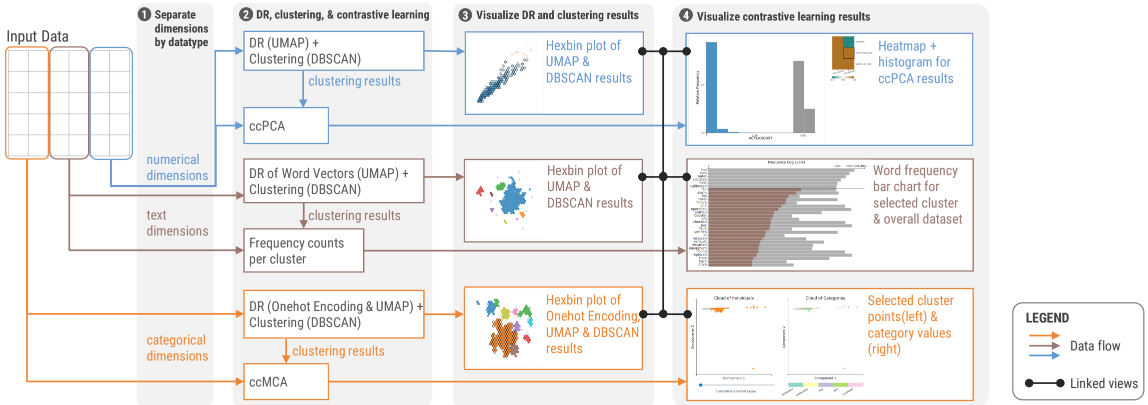

Input Data
1 Separate dimensions by datatype
2 DR, clustering, & contrastive learning
3 Visualize DR and clustering results
4 Visualize contrastive learning results
DR (UMAP) + Clustering (DBSCAN)
clustering results
ccPCA
numerical dimensions
DR of Word Vectors (UMAP) + Clustering (DBSCAN)
clustering results
text dimensions
Frequency counts per cluster
DR (Onehot Encoding & UMAP) + Clustering (DBSCAN)
clustering results
categorical dimensions
ccMCA
Hexbin plot of UMAP & DBSCAN results
Hexbin plot of UMAP & DBSCAN results
Hexbin plot of Onehot Encoding UMAP & DBSCAN results
Heatmap + histogram for ccPCA results
Word frequency bar chart for selected cluster & overall dataset
Selected cluster points(left) & category values (right)
LEGEND
Data flow
Linked views

Figure 1: Data processing pipeline for individual views based on the category of data dimensions (categorical, text, and numerical). The figure also shows which views are linked via selection and filtering interactions.

these elements can be classified based on their datatype as numerical, categorical, and text. They also propose a maintenance management workflow with six steps: (1) analyzing the work order, (2) selecting and prioritizing work orders, (3) planning equipment, resources, and labor, (4) scheduling the tasks involved, (5) executing the tasks, and (6) completing and documenting the tasks performed. Our goal is to aid the user—assumed to be a planning engineer or an analyst—in the execution of Steps 1–3. Depending on the scenario, this may require accurate identification of the maintenance task involved, using maintenance logs to anticipate component failure, or correcting work orders with misdiagnosed problems or misidentified tasks.

We thus infer that a system that uses maintenance log data to aid maintenance planning and management needs to be robust to different datatypes, supports visual analysis of data at scale, and helps the user characterize and label parts of the data based on their domain knowledge. The system requirements are:

R1 **Robustness to Datatype:** The system should accommodate all three types of data commonly required for the analysis of maintenance logs, i.e., numerical, categorical, and text data. Given the inherent difference between the datatypes, an appropriate analysis approach is needed for each.

R2 **Scalability:** Maintenance log data in an organization can vary from a few thousand records to hundreds of thousands of records, depending on the organization size. With each record consisting of several dimensions of mixed datatypes, the system needs to be robust to different data scales.

R3 **Data Subset Identification:** When visualizing large-scale data with heterogeneous dimensions, it is not optimal or practical to start by examining individual data points. It is more important and efficient to be able to identify subsets comprising data points that are closely related to each other. This may mean that all data points in an identified subset have common attributes, or that they may be related to each other based on their values along multiple dimensions. With different dimensions composed of different datatypes, the system should allow subset identification approaches suitable across datatypes.

R4 **Data Subset Characterization:** Analyzing maintenance logs requires not only the identification of patterns/subsets within the data, but also their *characterization*, or what separates them. For instance, a problem common to a group of machines could be characterized by all machines being similar (e.g., lathes), or requiring replacement of the same component, or of components supplied by the same vendor. Identifying such common characteristics become more difficult as the relationship shared by a subset of maintenance logs becomes more complex. Thus, the system should provide effective analysis support to characterize the subsets from many dimensions.

R5 **Extensibility:** Different organizations may choose to log information about their maintenance activity in different forms and granularities. The only aspect that may be common across these datasets is that they are likely to be multidimensional and heterogeneous. The system should be extensible to a different dataset with minimal effort, and not be overly dependent on any one specific dataset's attributes or format.

4 Data Processing & Visualization

Based on the requirements identified in Sect. 3, it is clear that the three types of data common to machine maintenance logs—numerical, categorical, and text—need to be processed appropriately and visualized using approaches that are robust to changes in the data scale. In this section, we describe the data processing approaches and visualization designs that address the identified requirements[1] .

4.1 Workflow

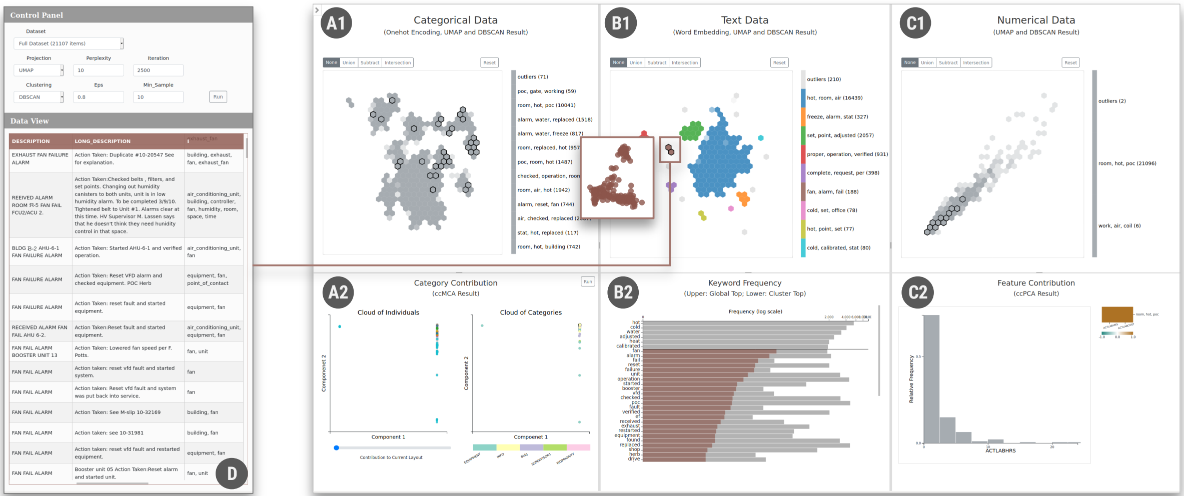
In Sect. 2, we see that visualizing heterogeneous data is advantageous as it allows the user to draw inferences based on context from different data dimensions. We also see that the issues of scale and dimensionality make it challenging for such observations and inferences to be drawn. Both issues are addressed by using clustering techniques to form subsets within the data (requirement **R3**). These can then be visually and interactively explored to understand the relationship between the data points that make up the subset.

To aggregate the techniques mentioned above, we model our data processing and visualization workflow as a pipeline with six steps: **Step 1**: grouping the data dimensions together based on their datatype (Fig. 1 stage 1); **Step 2**: performing DR for numerical, categorical, and text data separately and obtaining a 2D projection for each (Fig. 1 stage 2); **Step 3**: clustering the 2D data to form subsets (Fig. 1 stage 2); **Step 4**: visualizing the 2D projection and clustering results to provide scalable overviews of the dataset (Fig. 1 stage 3 and Fig. 2 A1, B1, C1); **Step 5**: characterizing the clusters separately for each datatype using contrastive learning or statistical methods (Fig. 1 stage 2); **Step 6**: cluster characterization for each datatype with an appropriate visualization (Fig. 1 stage 4 and Fig. 2 A2, B2, C2). Each step is detailed in the rest of this section.

4.2 Identifying Subsets in Heterogeneous Data

DR (step 2) and clustering (step 3) are two essential data processing steps to identify subsets in the data. Informed by our review in Sect. 2.3, we choose UMAP [36] to project the data to a lower-dimensional space. By using a nonlinear DR method such as UMAP, we can effectively extract similar records from high-dimensional

[1] The source code is available at https://github.com/Xiaoyu1993/Machine-Maintenance-Log-Analysis.



Control Panel

Dataset: Full Dataset (21107 items)

Projection: UMAP | Perplexity: 10 | Iteration: 2500

Clustering: DBSCAN | Eps: 0.8 | Min_Sample: 10 | Run

Data View

DESCRIPTION	LONG_DESCRIPTION	exhaust_fan
EXHAUST FAN FAILURE ALARM	Action Taken: Duplicate #10-20547 See for explanation.	building, exhaust, fan, exhaust_fan
REEIVED ALARM ROOM ℝ-5 FAN FAIL FCU2/ACU 2.	Action Taken:Checked belts , filters, and set points. Changing out humidity canisters to both units, unit is in low humidity alarm. To be completed 3/9/10. Tightened belt to Unit #1. Alarms clear at this time. HV Supervisor M. Lassen says that he doesn't think they need hunidity control in that space.	air_conditioning_unit, building, controller, fan, humidity, room, space, time
BLDG 𝔹-2 AHU-6-1 FAN FAILURE ALARM	Action Taken: Started AHU-6-1 and verified operation.	air_conditioning_unit, fan
FAN FAILURE ALARM	Action Taken: Reset VFD alarm and checked equipment. POC Herb	equipment, fan, point_of_contact
FAN FAILURE ALARM	Action Taken: reset fault and started equipment.	equipment, fan
RECEIVED ALARM FAN FAIL AHU 6-2.	Action Taken:Reset fault and started equipment.	air_conditioning_unit, equipment, fan
FAN FAIL ALARM BOOSTER UNIT 13	Action Taken: Lowered fan speed per F. Potts.	fan, unit
FAN FAIL ALARM	Action taken: reset vfd fault and started system.	fan
FAN FAIL ALARM	Action taken: Reset vfd fault and system was put back into service.	fan
FAN FAIL ALARM	Action Taken: See M-slip 10-32169	building, fan
FAN FAIL ALARM	Action taken: see 10-31981	building, fan
FAN FAIL ALARM	Action taken: reset vfd fault and restarted equipment.	equipment, fan
FAN FAIL ALARM	Booster unit 05 Action Taken:Reset alarm and started unit.	fan, unit

Figure 2: Dashboard interface showing projected views of categorical components (A1), text components (B1), and numerical components (C1) of the dataset using the DR algorithm UMAP. Each projected view is clustered using a chosen clustering algorithm (DBSCAN in the example above). Each projected view is supported by an additional view that is used to characterize a chosen cluster in that view. For the categorical data view, ccMCA (A2) is used to show the selected cluster's separation and the attribute values that contribute to it. A text frequency chart (B2) contrasts the text that occurs most frequently in the selected cluster against the overall text frequency in the dataset. Finally, ccPCA is used to display a heatmap of cluster vs. data dimensions and a histogram showing the value distribution of a selected numerical dimension against the rest of that data (C2). Raw data for any chosen cluster can be viewed using a slide-out tabular view (D). Linking across views A1, B1, and C1 shows the distribution of data clustered in the active view (in color) across the other two views (grayscale).

maintenance log data. High-dimensional representations are obtained for the categorical data with one-hot encoding [23], and for text with word embeddings (see Sect. 4.4.1) before the DR step. The 2D projection of the data can then be clustered using any approach.

We choose DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [16] as it uses a density-based approach that is more suitable for data that may have outliers (e.g., an unusual machine breakdown or repair). Combined with our visualization approach, this technique is more suitable for our case as the analyst can probe into individual records in the case of outliers, and can also examine larger clusters using the linked views. By separating the data dimensions based on datatype, we ensure that our approach is robust to datasets with different dimensions with mixed datatypes (addressing requirement **R1**). This approach of dimension grouping by datatype, DR & clustering to find subsets, and characterizing based on contrastive learning and text frequency comparison makes our approach extensible to most heterogeneous datasets (**R5**).

4.3 Scalable Overview

To show an overview of the DR and clustering results (step 4), we use a hexbin plot [10] for each datatype in the dashboard visualization shown in Fig. 2, i.e., categorical (Fig. 2A1), text (B1), and numerical (C1) dimensions. The hexbin plot is robust to different data scales (requirement **R2**) in that its rendering speed is not significantly impacted by data size or screen resolution. Instead of using a linear color scale typical to hexbin plots, we use a different hue for each cluster and map data density to color intensity within every cluster.

We also preserve the conventional DR representation, i.e., a scatterplot with each data object shown by a dot. We adopt Lindstrom's [31] Level of Detail (LOD) rendering and allow users to switch between these two plots or change the granularity of hexagonal bins by simply zooming in or out of the area they are interested in. Thus, only a small part of the scatterplot needs to be rendered when the users zoom in close. Finally, users can choose to examine the data objects in detail by perusing the slide-out tabular view (Fig. 2D) or by hovering over the dots.

Note that at any point, only one of the three clustered views (A1, B1, or C1 in Fig. 2) can be active. The active view is indicated by its clusters highlighted with a categorical color palette. The remaining views are monochromatic/greyscale to prevent the user from mistakenly assuming that a cluster of one color (e.g., blue) in one view corresponds to a cluster of the same color in another view.

4.4 Characterizing clusters

Characterizing a cluster or subset in the data (requirement **R4**) requires the determination of how the cluster is different from the rest of the data. Different datatypes necessitate different contrastive analysis techniques. We discuss the techniques we use to characterize clusters for text, numerical, and categorical data in this subsection.

4.4.1 Text Dimensions

Detailed text descriptions of problems, symptoms, and solutions, form perhaps the richest component of maintenance log data. They are also rife with inconsistencies, typographical errors, or the use of non-standard shorthand that is endemic to that particular organization. Text descriptions are also often supplemented by "tags"—standardized phrases that label the descriptions to identify the problems, items, and solutions. These tags are typically assigned partly based on the knowledge of the user who tags the descriptive text, and partly using machine learning approaches [43,45].

In order to group the data based on text dimensions, the *meaning* of the text needs to be considered instead of specific keywords that may vary across technical personnel. A more consistent semantic representation would focus on the meaning of the text rather than its form, such that synonyms and related terms are grouped closely. To achieve this, we use word embeddings, which are vector representations of words that take into account their semantic relationships [53]. Words such as "warm" and "hot" can thus be translated to vectors that are close to each other, but distant from a vector representing a word different in meaning, such as "telephone". We create high-dimensional vector representations for the descriptive text by summing and normalizing words in the text. We then use a suitable DR technique (UMAP) to obtain 2D projections of the vectors, and cluster them using DBSCAN (Fig. 3).

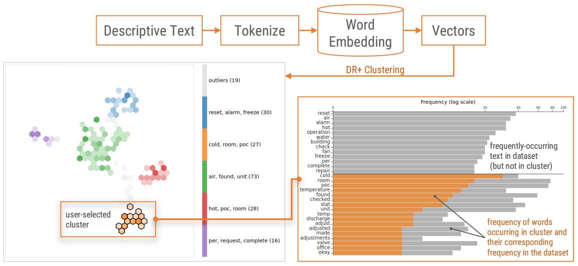


Figure 3: Text processing steps and visualization of the related information. The text frequency chart shows frequencies of text occurring in the selected cluster and contrasts it with both, the frequencies of the corresponding text in the overall dataset as well as the text that is most frequently-occurring in the dataset but not in the selected cluster.

Each cluster represents a collection of descriptions. To characterize a given cluster, we overlay a frequency plot of the most common terms occurring in the cluster on a frequency plot of terms occurring in the overall dataset (Fig. 3 right). Contrasting the most frequent terms of both plots helps the user identify defining characteristics of the cluster. For instance, by examining the frequency plots of Cluster 1 in Fig. 3, we can surmise that the cluster represents maintenance logs of ventilation systems related to lower room temperatures, commonly remedied by adjusting certain valves. The user can examine the raw data related to any cluster using the slide-out tabular view (Fig. 2D) to further gain insight into the cluster and characterize it (requirement **R4**). The cluster labels are editable. For instance, the user can replace the default cluster name with a more descriptive phrase "Lower temperature adjustment".

4.4.2 Numerical Dimensions

As Brundage et al. [7] illustrated with various maintenance key performance indicators (KPIs), measures such as the number of problems/breakdowns, time between failures, and time taken to repair can be used to quantify the role of other performance indicators, such as machine type, problem severity, and technician skill. Other parameters such as cost can be derived from these factors. To understand how those parameters contribute to the separation of clusters for numerical data, we adopt a method called ccPCA [20]. We briefly describe ccPCA and its application to our system. Notations used in the following sections are summarized in Table 1.

Introduction to cPCA. cPCA aims to reveal enriched patterns in a target matrix $\mathbf{X}_T$ relative to a background matrix $\mathbf{X}_B$. To do so, cPCA finds directions (called contrastive principal components, cPCs) that maximally preserve the variation in $\mathbf{X}_T$ while simultaneously minimizing the variation in $\mathbf{X}_B$. This can be achieved by performing EVD on $(\mathbf{C}_T - \alpha\mathbf{C}_B)$ where $\mathbf{C}_T$ and $\mathbf{C}_B$ are covariance matrices of $\mathbf{X}_T$ and $\mathbf{X}_B$, respectively. α $(0 \leq \alpha \leq \infty)$ is a hyperparameter, called a contrast parameter, which controls the trade-off between having high target variance and low background variance. When $\alpha = 0$, the resultant cPCs only maximize the variance of $\mathbf{X}_T$ (i.e., the same with using ordinary PCA). As α increases, cPCs place greater emphasis on directions that reduce the variance of $\mathbf{X}_B$.

Introduction to ccPCA. In order to characterize clusters, ccPCA utilizes cPCA as its base. Let $\mathbf{X}_E$, $\mathbf{X}_K$, and $\mathbf{X}_R$ be matrices of the entire dataset, a target cluster selected from the entire dataset, and the rest of the data points, respectively. ccPCA enhances the original cPCA by using $\mathbf{X}_E$ as a target matrix and $\mathbf{X}_R$ as a background matrix, instead of using $\mathbf{X}_K$ and $\mathbf{X}_R$ as target and background matrices, respectively. With the automatic selection of a contrast parameter [20], ccPCA finds the directions that preserve both the variety and separation between a target cluster and others. These directions are difficult to find with the original cPCA (see the work by Fujiwara et al. [20] for details). By referring to feature contributions (called contrastive principal component loadings or cPC loadings) to the directions, we can obtain the information of which numerical features contribute to the uniqueness of a target cluster relative to others.

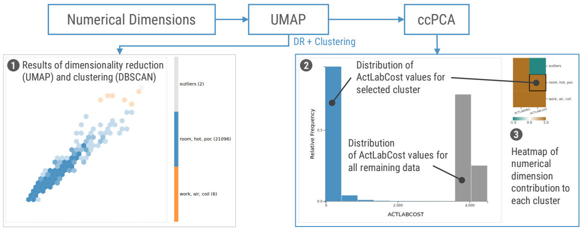

Numerical Dimensions
UMAP
ccPCA
DR + Clustering
1
Results of dimensionality reduction (UMAP) and clustering (DBSCAN)
outliers (2)
room, hot, poc (21096)
work, air, coil (6)
2
Distribution of ActLabCost values for selected cluster
Distribution of ActLabCost values for all remaining data
Relative Frequency
0.0
0.5
0
2,000
4,000
ACTLABCOST
outliers
room, hot, poc
work, air, coil
ACTLABHRS
ACTLABCOST
-1.0
0.0
1.0
3
Heatmap of numerical dimension contribution to each cluster

Figure 4: Projection and visualization of numerical data along with the ccPCA view show that the selected cluster has a lower labor cost than the rest of the data.

Table 1: Summary of notations.

$\mathbf{X}_T$, $\mathbf{X}_B$	target, background matrices
$\mathbf{X}_E$, $\mathbf{X}_K$, $\mathbf{X}_R$	matrices of the entire, target cluster, rest data points
$\mathbf{C}_T$, $\mathbf{C}_B$, $\mathbf{C}_E$, $\mathbf{C}_R$	covariance matrices of $\mathbf{X}_T$, $\mathbf{X}_B$, $\mathbf{X}_E$, $\mathbf{X}_R$
$\mathbf{G}_T$, $\mathbf{G}_B$, $\mathbf{G}_T$, $\mathbf{G}_R$	disjunctive matrices of $\mathbf{X}_T$, $\mathbf{X}_B$, $\mathbf{X}_E$, $\mathbf{X}_R$
$\mathbf{Z}_T$, $\mathbf{Z}_B$, $\mathbf{Z}_E$, $\mathbf{Z}_R$	probability matrices of $\mathbf{G}_T$, $\mathbf{G}_B$, $\mathbf{G}_T$, $\mathbf{G}_R$
$\mathbf{B}_T$, $\mathbf{B}_B$, $\mathbf{B}_E$, $\mathbf{B}_R$	Burt matrices of $\mathbf{X}_T$, $\mathbf{X}_B$, $\mathbf{X}_E$, $\mathbf{X}_R$
α	contrast parameter

Visualization. ccPCA provides how strongly each dimension contributes (positively or negatively) to each cluster's contrast with the rest of the data. This contribution is shown as a heatmap (Fig. 4(3)) that indicates the magnitude and direction of the contribution of the numerical dimensions to each cluster with a blue-green-to-brown diverging colormap. By selecting a cell in the heatmap, Fig. 4(2) shows histograms of the corresponding dimension's value distributions of the selected cluster and the rest of the data with the cluster color and gray color, respectively. Based on Fig. 4(2), we can infer that the numerical dimension "actual labor cost" (`ActLabCost`) contributes strongly to Cluster 0's contrast against the rest of the data, and the histograms show that the `ActLabCost` values for the selected cluster are much lower than the rest of the data. The user can further investigate this cluster by selecting it in the DR view (Fig. 4-1) to examine the corresponding data distribution in the text and categorical dimension views as described in Sect. 4.3, or examine the cluster in detail using the tabular view (Fig. 2(D)). Note that Fig. 4 shows only two numerical dimensions due to the dataset we analyze; however, as demonstrated in [20], the combination of using DR, clustering, and ccPCA is useful in identifying and characterizing subsets within high-dimensional numerical data.

4.4.3 Categorical Dimensions

We cannot use ccPCA—which requires numerical or binary data—to characterize categorial data (**R4**). Thus, we introduce a new contrastive learning method, called contrasting clusters in multiple correspondence analysis (ccMCA) by extending multiple correspondence analysis (MCA). Table 2 compares the related methods.

Multiple Correspondence Analysis (MCA) Here, we provide a brief introduction to MCA (refer to [30] for details). MCA can be considered as PCA for categorical data. That is, MCA learns a lower-dimensional representation from high-dimensional categorical data as it maximally preserves the variance of the data. The issue of PCA when applying to categorical data is that PCA handles each category in the data as a numerical value and, as a result, it unnecessarily ranks the categories (e.g., red: 0, green: 1, blue: 2).

To avoid this, MCA first converts an input matrix $\mathbf{X}_T$ of categorical data into a disjunctive matrix $\mathbf{G}_T$ (or disjunctive table) by applying one-hot encoding to each categorical dimension. For example, when $\mathbf{X}_T$ consists of two columns (or often called questions) of "color" and "shape" and each has categories (i.e., categorical answers) of {"red", "green", "blue"} and {"circle", "rectangle"}, $\mathbf{G}_T$ will have five columns of "red", "green", "blue", "circle", and "rectangle" and each of the matrix elements will be either 0 or 1. Afterward, by dividing each cell in $\mathbf{G}_T$ with a total of $\mathbf{G}_T$, we obtain

Zhang, Xiaoyu; Fujiwara, Takanori; Chandrasegaran, Senthil; Brundage, Michael; Sexton, Thurston; Dima, Alden; Ma, Kwan-Liu. "A Visual Analytics Approach for the Diagnosis of Heterogeneous and Multidimensional Machine Maintenance Data." Presented at 14th IEEE Pacific Visualization Symposium (PacificVis 2021). April 19, 2021 - April 21, 2021.

Table 2: Comparison of representation learning methods. ccMCA is a new method we introduce in this paper.

data type	method	purpose	solution
numerical, binary	PCA	preserving the variance of $\mathbf{X}_T$	EVD on $\mathbf{C}_T$
	cPCA	identifying enriched patterns in $\mathbf{X}_T$	EVD on $(\mathbf{C}_T - \alpha\mathbf{C}_B)$
	ccPCA	characterizing a cluster $\mathbf{X}_K$	EVD on $(\mathbf{C}_E - \alpha\mathbf{C}_R)$
categorical, binary	MCA	preserving the variance of $\mathbf{X}_T$	EVD on $\mathbf{B}_T$
	cMCA	identifying enriched patterns in $\mathbf{X}_T$	EVD on $(\mathbf{B}_T - \alpha\mathbf{B}_B)$
	ccMCA	characterizing a cluster $\mathbf{X}_K$	EVD on $(\mathbf{B}_E - \alpha\mathbf{B}_R)$

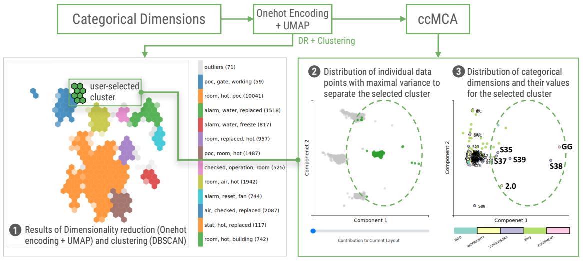


Figure 5: Projection of categorical data (1), with the ccMCA view showing the separation of the selected cluster (2), and its corresponding category distribution (3). The categories of "GG" in "EQUIPMENT", 2.0 in "WOPRIORITY" (work priority), and S35–S39 in "SUPERVISOR1" are most likely to be the characterization of this cluster.

a probability matrix (or correspondence matrix) $\mathbf{Z}_T$. This probability matrix corresponds to an input feature matrix for PCA. Similar to PCA, we apply normalization to $\mathbf{Z}_T$. With the normalized $\mathbf{Z}_T$, we can obtain a Burt matrix, $\mathbf{B}_T$, with $\mathbf{B}_T = \mathbf{Z}_T^\top \mathbf{Z}_T$. $\mathbf{B}_T$ corresponds to a covariance matrix used in PCA (note: in PCA, a covariance matrix of $\mathbf{X}_T$ can be obtained with $\mathbf{C}_T = \mathbf{X}_T^\top \mathbf{X}_T$). Thus, as PCA obtains principal components by performing eigenvalue decomposition (EVD) on $\mathbf{C}_T$, MCA obtains the principal directions by performing EVD on $\mathbf{B}_T$ to preserve the variance of $\mathbf{G}_T$.

Contrastive MCA (cMCA) Now, we introduce contrastive version of MCA (cMCA) [21] and enhance cMCA to ccMCA in the next subsection. As described above, MCA and PCA fundamentally share the same idea of finding the best directions to preserve the variance by using EVD on a covariance matrix. Therefore, we can extend MCA to cMCA by employing the same idea with cPCA.

Extension from MCA to cMCA. As described in Sect. 4.4.2, the only difference between PCA and cPCA is that while PCA directly performs EVD on a target covariance matrix $\mathbf{C}_T$, cPCA takes a subtraction of target and background covariance matrices with a contrast parameter (i.e., $\mathbf{C}_T - \alpha\mathbf{C}_B$) and then performs EVD on it. To reveal enriched patterns in a target matrix of categorical values, we can use the same idea that we use with cPCA and apply it to MCA. As stated in Sect. 4.4.3, in MCA, a Burt matrix $\mathbf{B}_T$ contains similar information with a covariance matrix $\mathbf{C}_T$ in PCA. Therefore, we can obtain contrastive directions by computing $\mathbf{B}_T - \alpha\mathbf{B}_B$, where $\mathbf{B}_T$ and $\mathbf{B}_B$ are target and background Burt matrices, and then performing EVD on $(\mathbf{B}_T - \alpha\mathbf{B}_B)$. Here, α $(0 \leq \alpha \leq \infty)$ is also a contrast parameter and has the same role as cPCA.

Contrasting Clusters in MCA (ccMCA) For the cluster characterization, we enhance cMCA to ccMCA. Here, we apply the similar idea of the extension from cPCA to ccPCA.

Extension from cMCA to ccMCA. cMCA can be enhanced to ccMCA by using $\mathbf{X}_E$ and $\mathbf{X}_R$ as input target and background matrices. Since the directions identified by ccMCA differ based on the contrast parameter α, we also provide the automatic selection method of α by employing the same method introduced by Fujiwara et al. [20], which utilizes the histogram intersection for its optimization. Fig. 5(2) shows the ccMCA result when selecting the green cluster from Fig. 5(1) as a target cluster. The green points are clearly separated from others while keeping a high variance.

One ccMCA's major and different challenge from ccPCA is that how we inform the feature contributions. ccMCA also provides contributions (or loadings) of each dimension (i.e., category) of $\mathbf{G}_T$ with $\mathbf{w}_i = \sqrt{\lambda_i}\mathbf{v}_i$ where $\mathbf{w}_i$ is feature contributions to the i-th principal direction, λ_i is the i-th top eigenvalue generated via EVD, and $\mathbf{v}_i$ is the corresponding eigenvector. Because EVD is performed on Burt matrices of $\mathbf{G}_T$ and $\mathbf{G}_B$, which are obtained by applying one-hot encoding to $\mathbf{X}_T$ and $\mathbf{X}_B$, $\mathbf{w}_i$ shows a contribution for each category (e.g., "red", "green", and "blue") but not for each question (e.g., "color"). Therefore, the number of dimensions of $\mathbf{w}_i$ can be easily overwhelmed. For example, when there are 6 questions and 5 categories for each question, the number of dimensions in $\mathbf{w}_i$ becomes 30. Also, as each data point's position in a ccMCA projection (e.g., Fig. 5(2)) reflects a compound of contributions, looking at each contribution may not be sufficient to understand the association between the projection and contributions. For instance, even when one category may have a strong contribution to the positive direction of the first axis (x-axis in Fig. 5(2)), this does not ensure that data points with large positive x-coordinates have answered the corresponding category because, at the same time, many other categories may have a weak contribution to the positive direction.

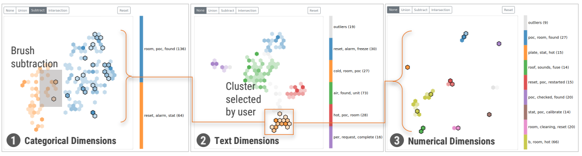


Figure 6: Linking among the projected views of categorical, text, and numerical data allows the user to explore the data clusters from the perspective of datatypes. For instance, selecting Cluster "cold, room, poc" from the projected and clustered view of the text dimensions (2) highlights the distribution of the same points in the other two views (1) & (3). We can see some correlation between the selected cluster and Cluster "room, poc, found" in the categorical dimension view. The brush and Boolean subtraction tools can be used to refine the selection and further reveal the correlation between the two clusters.

To address this issue, similar to MCA, we provide the *principal cloud of categories* (or *column principal coordinates*), as shown in Fig. 5(3). In MCA, the principal cloud of categories (PCC) is used to grasp which categories each data point likely has answered by comparing the positions of data points in an MCA projection (or the *principal cloud of individuals*, PCI) and categories in PCC. When a data point in PCI is placed at a close position with certain categories in PCC, this data point tends to have these categories as its answers. We can also perform the same analysis above for ccMCA.

In MCA, PCC $\mathbf{Y}_T^{\text{col}}$ is usually obtained by taking a product of a diagonal matrix $\mathbf{D}_T$ of the sum for each column of $\mathbf{G}_T$ and the top-k eigenvectors $\mathbf{W}_T$ obtained by EVD (i.e., $\mathbf{Y}_T^{\text{col}} = \mathbf{D}_T\mathbf{W}_T^\top$). However, because ccMCA performs EVD on $(\mathbf{B}_T - \alpha\mathbf{B}_B)$ and the result is influenced by $\mathbf{X}_B$ as well, we cannot compute PCC in the above manner. Instead, we use MCA's *translation formula* from PCI to PCC [30]. The translation from PCI to PCC can be performed with:

$$\mathbf{Y}_{\text{col}} = \mathbf{D}_T^{-1}\mathbf{Z}_T^\top \mathbf{Y}_T^{\text{row}} \text{diag}(\boldsymbol{\lambda})^{-1/2} \tag{1}$$

where $\mathbf{Y}_T^{\text{row}}$ is PCI of a target matrix and $\boldsymbol{\lambda}$ is a vector of the top-k eigenvalues. An example of the resultant PCC is shown in Fig. 5(3). By referring to Fig. 5(2) and (3), the analyst can characterize a selected cluster by understanding which categories are highly associated with the uniqueness of the cluster.

4.5 Linking and Interactions

The visualizations across all six panels of the dashboard and tabular view are fully linked, and support brushing and direct selection (of a bin/cluster). Users can select, say, a cluster of interest in one of the projected 2D views (A1, B1, or C1 in Fig. 2) and observe the distribution of the cluster in the remaining two views. Each projected view is supported by a cluster characterization view (A2, B2, and C2 in Fig. 2). When a cluster is selected from one of the projected

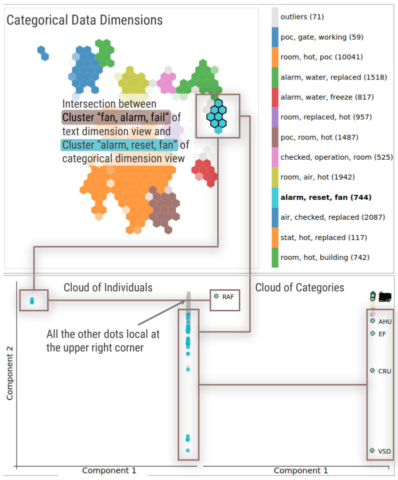

Categorical Data Dimensions
outliers (71)
poc, gate, working (59)
room, hot, poc (10041)
alarm, water, replaced (1518)
alarm, water, freeze (817)
room, replaced, hot (957)
poc, room, hot (1487)
checked, operation, room (525)
room, air, hot (1942)
alarm, reset, fan (744)
air, checked, replaced (2087)
stat, hot, replaced (117)
room, hot, building (742)
Intersection between Cluster "fan, alarm, fail" of text dimension view and Cluster "alarm, reset, fan" of categorical dimension view
Cloud of Individuals
Cloud of Categories
RAF
All the other dots local at the upper right corner
AHU
EF
CRU
VSD
EQUIPMENT
INFO
Bldg
WOPRIORITY
SUPERVISOR1
Component 2
Component 1
Component 1

Figure 7: Characterizing a customized cluster (see Sect. 6-Cluster Characterization). Dots in the cloud of categories that share similar locations to the colored dots in the cloud of individuals reveal equipment types that contribute more to the separation of this cluster.

views, all three characterization views update to show the results of that cluster's characterization analysis based on categorical (A2), text (B2), and numerical (C2) data dimensions. The tabular view also updates to show the attributes of the data in the selected cluster.

The linked views update in a similar manner even if—instead of selecting a cluster—the user selects, say, a single hexbin, or brushes across multiple hexbins. Boolean operations such as the union, intersection, and difference are also supported for more sophisticated selections of data across the three projected views. For instance, the user can intersect multiple clusters across different views to find points common across clusters, or combine the clusters by the union.

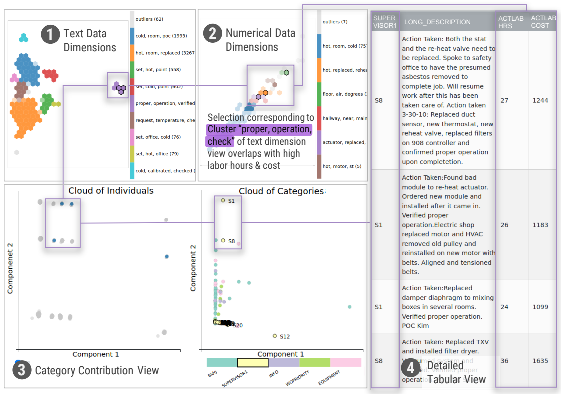
Fig. 6 shows an example of interactive linking. The user selects the cluster labeled "cold, room, poc" in panel 2 (projected view of text). This highlights hexbins in the other two views that correspond to this cluster. In the example shown, most of the data points overlap with Cluster "room, poc, found" in panel 1 (categorical dimensions), indicating a correlation between these two clusters. To better observe the overlapping points, the user subtracts the two outliers in panel 1 by brushing them out, and checks the supplementary views. Panel 3 shows the points distributed across clusters, indicating no correlation between the selected clusters along their numerical dimensions.

5 Implementation

The dashboard visualization is implemented as a web framework with a Flask server at the back end. The separation of numerical, categorical, and text dimensions is currently performed manually. We compute dimensionality reduction and clustering at the back end for each of these three groups of dimensions and visualize the results by creating an interactive web-based dashboard application. We use HTML/JavaScript for the front end using Bootstrap and React libraries, and D3 [4] to create interactive visualizations.

We use the Scikit-Learn [38] machine learning library for most of the dimensionality reduction and clustering algorithms, except for UMAP and ccPCA, for which we use implementations by McInnes et al. [36] and Fujiwara et al. [20] respectively. We use our own implementations of MCA and ccMCA for DR and contrastive learning for categorical dimensions. For the text dimensions, we use the Natural Language Toolkit (NLTK) [33] for the text processing, ConceptNet Numberbatch [48] as the word embedding to vectorize the text, and Gensim [41] to perform the word-vector lookup.

We tokenize the descriptive text and tags, and remove stop words. Vector representations of words in the remaining text are retrieved using the word embedding, normalized, and added to obtain a single vector representing the unstructured text component of each data point. While ConceptNet Numberbatch contains a fairly large vocabulary of over 500,000 words, there may be domain- or organization-specific terms used in maintenance logs that are not present in the word embedding. In our current implementation, we discard these terms on the assumption that enough of the meaning is captured in the rest of the text for clustering data. However, in future iterations, we plan to update word embeddings using vocabulary from technical manuals and organizational documentation.



SUPER VISOR1	LONG_DESCRIPTION	ACTLAB HRS	ACTLAB COST
S8	Action Taken: Both the stat and the re-heat valve need to be replaced. Spoke to safety office to have the presumed asbestos removed to complete job. Will resume work after this has been taken care of. Action taken 3-30-10: Replaced duct sensor, new thermostat, new reheat valve, replaced filters on 908 controller and confirmed proper operation upon completetion.	27	1244
S1	Action Taken:Found bad module to re-heat actuator. Ordered new module and installed after it came in. Verified proper operation.Electric shop replaced motor and HVAC removed old pulley and reinstalled on new motor with belts. Aligned and tensioned belts.	26	1183
S1	Action Taken:Replaced damper diaphragm to mixing boxes in several rooms. Verified proper operation. POC Kim	24	1099
S8	Action Taken: Replaced TXV and installed filter dryer. [illegible] operat [illegible]	36	1635

Figure 8: Examining a subset of the original data characterized by temperature-related complaints. The selected purple cluster is—using the category contribution view (3)—shown to be related to high costs associated with two supervisors (see Sect. 6).

6 Use Case Scenario

We illustrate the use of our system using maintenance log data of HVAC systems used in multiple office buildings of an organization. We focus on tasks where the user would analyze the data for patterns and trends to better allocate resources, assign technicians, and schedule future work. Such tasks are an important part of the maintenance cycle, as described in Brundage et al. [8]. The maintenance logs consist of over 21,000 records collected over ten years and contain multiple dimensions of categorical, text, and numerical data. For the purpose of this use-case scenario, we select dimensions of the data that have the least number of missing values. The dataset is grouped by the following sets of dimensions.

The first group involves the categorical dimensions of (1) the `building number` where a complaint on the HVAC system was recorded, (2) `equipment type` of the HVAC subsystem or machine, (3) work order `priority`, (4) system/complaint `location` (building number + floor + room), (5) the index of the `supervisor` in charge of the systems at the time of logging the problem/solution.

Most of the numerical dimensions in the data involve dates and times of logging and are not accurate or consistent enough to compute meaningful timespans. The second group thus involves the two remaining numerical dimensions of (1) `actual labor hours` incurred, and (2) `actual labor cost` incurred.

The third group consists of the text dimensions of (1) `long description` or a description of the problem or complaint that needed addressing, (2) `description` or a small set of keywords highlighting the important aspects of the problem, and (3) a set of multiple `tags` assigned to each maintenance record. The text fields were cleaned up to remove extraneous characters (e.g., HTML tags, symbols, URLs, etc.), remove punctuation, normalize whitespace sequences, and correct typographical and Unicode errors.

Our scenario involves Alice, a maintenance supervisor responsible for the smooth running of HVAC systems across the organization. Alice uses our prototype to examine the dataset and identify patterns in the logs to identify potential issues and plan preventive maintenance. One of her initiatives has been to try and allocate manpower for recurring or preventable maintenance problems.

Overview. Alice loads all three data groups discussed above into the prototype to get an overview of the data after DR and clustering. She looks over the default tags assigned to each cluster and notices such commonly-occurring terms as *"room"*, *"air"*, *"hot"*, and *"cold"*. The largest cluster in panel B1 (Fig. 2) showing text dimensions appears to contain complaints related to room temperature, with the tags *"too hot"* and *"too cold"* being the most common. From experience, she figures these represent the most typical complaints about HVAC systems in offices. The top keywords in the frequency plot (B2 in Fig. 2) confirm her hunch.

Looking over at the numerical data projection (C1 in Fig. 2), Alice notices that it appears to be linearly correlated. Examining the heatmap in the feature contribution panel (C2 in Fig. 2) confirms this observation as she finds that the *"actual labor hours"* do indeed correlate with *"actual labor cost"*. She makes a mental note to refer to the correlation to filter the data by time or cost in her analysis.

Cluster Characterization. Apart from the clusters related to the temperature problems, Alice notices a unique brown cluster in the text dimension views tagged as "*fan, alarm, fail*" (B1 in Fig. 2) and decides to take a closer look at it. From panel B2, she finds that the top keywords in this cluster—*fan, alarm, fail, reset, repair*—are significantly different from those in the rest of the dataset. She also finds that the cluster overlaps with all clusters in the categorical view (A1 in Fig. 2) that have the above three terms as one of their main tags. The highest overlap in the categorical data view is with a cyan cluster (see Fig. 7) tagged with *"alarm"*, *"reset"*, and *"fan"*.

She uses the Boolean operator to separate the intersection between these two clusters. From the category contribution panel (Fig. 7), she notices that several types of equipment including *"RAF"* (Return Air Fan), *"EF"* (Exhaust Fan), *"VSD"* (Variable Speed Drive), *"CRU"* (Customer Replaceable Unit), and *"AHU"* (Air Handling Unit) contribute the most to this cluster. From her experience, she knows that the above equipment has always had relatively unreliable fans. Cross-checking with the numerical data panel, she realizes that the labor cost is relatively low for these problems, so she makes a note to have regular preventive maintenance done on the equipment.

Projection Interpretation. Now Alice decides to have the "too hot/cold" issue looked into further, and calls in an engineer to filter the dataset by these two tags and examine this filtered dataset separately. After loading the subset into the system, she notices a symmetry in the layout pattern of the text dimension view, about a horizontal axis. The clusters in the upper half of the projection all contain the keyword *"cold"* while those on the lower half contain *"hot"*. She infers that the vertical direction in the projected space relates to temperature, and becomes interested in the clusters located in the middle, especially the solitary purple cluster with tags *"proper, operation, verified"* (Fig. 8-1). She notices a significant variation of labor cost and hours in this cluster (Fig. 8-2). Selecting all points with a higher labor cost and hours, she learns from the updated keyword frequency plot that they correspond to the action *"replaced"*. From the category contribution panel, she finds that this part of the data is highly related to two supervisors *"S1"* and *"S8"* (Fig. 8-3). She confirms this observation by checking the tabular view (Fig. 8-4). She believes that the high cost may either a clerical mistake or an issue with the vendor supplying the parts. She decides to talk with these two supervisors to get to the bottom of the issue.

7 Expert Review

Our prototype was reviewed by three experts in machine maintenance analysis to determine its usefulness and throw light on the kinds of patterns or insights it might reveal to domain practitioners. The first expert (E1) was a data scientist from the industry who had developed approaches for extracting actionable information from maintenance data for over six years. The second expert (E2) was an industrial engineer specializing in model-based systems engineering methodologies. Finally, the third expert (E3) was a computer scientist from academia who worked on algorithm development and natural language processing for 25 years.

We conducted two pilot studies with co-authors of this work who are also domain experts in maintenance log analysis. This helped us simulate the remote setup and related logistics planned for the study, determine needs for—and forms of—tutorials and examples, and identify tasks that could be performed within existing constraints. Based on these, we designed a semi-structured, open-ended expert review. We followed the "pair analytics" paradigm [1] with one of the authors as the experimenter and the participant as the subject matter expert (SME). Pair analytics has been shown to be optimal for open-ended studies with SMEs, since it stimulates dialogue between the experimenter and the participant and explicates the participant's thought process, and reduces the burden of fluency with the application from the SME who is usually not a visual analytics expert [1]. The additional constraint of the pandemic, requiring a remote setup, further informed the decision to employ this paradigm.

The study used a video conference setup where the experimenter controlled the tool while the expert observed the visualizations and suggested filters, queries, and interactions via screen sharing. The experts perused a document explaining the views and functions of the prototype prior to the study, and were shown a 20-minute tutorial demonstrating the prototype at the start of the study. They were then asked to explore two datasets (30 mins each), one the HVAC dataset described in Sect. 6, and the other a subset of 17,000 records from the HVAC dataset involving temperature-related complaints.

We categorize our observations on the domain experts' remarks during the exploratory tasks and their feedback on the prototype into *functionality* of the prototype, *visual encoding*, and *interaction*.

Functionality. The tutorial and demonstration at the start of the study involved use cases and observations such as the one presented in Sect. 6. At the end of the demonstration, all three experts found our workflow to be *"highly reasonable"* (E2) and found the cases compelling. Yet, during the exploratory part of the study, they found it difficult to pin down the questions they could ask and answer of the data. For instance, E2 asked, *"What key question am I to answer?"* Based on the experts' questions about the visualizations, filters, and interactions during the exploratory study, we infer that the difficulty encountered by the experts was partly due to the relatively short time they spent with the interface and their unfamiliarity with the data.

Visual Encoding. Domain experts found the linked views to be intuitive and useful. E3 remarked, *"I like the ways that the panels are automatically updated with respect to the selections that (are) made. And being able to see the three types of data all together is good. Definitely a good idea to have them combined"*. However, they found it too abstract to separate the data dimensions into categorical, numerical, and text. As E1 explained, *"Looking at categorical, text and numerical data makes sense from a data perspective, but it's not necessarily the functional break down that makes sense."* Instead, they reported that they would have preferred a way of representing the data that allowed them to see the problems in a **functional** way, e.g., wherein the building, or wherein the machine a problem occurred, or what temporal patterns were observable in the data. E1 and E3 also found it a little confusing that the default cluster labels in the numerical and categorical panels still used keywords from the text component of the data. On the other hand, while they were able to characterize at least one of the clusters, none of them re-labeled the cluster(s). All three experts also found the characterization view for the categorical data difficult to understand. E1 said that they had *"a really hard time understanding this visualization"*. E3 noted that they had *"never seen the information displayed in this way with two side by side panels of the cloud of individuals and categories... it's a little non-specific as far as whether the dots that show up in the (cloud of) categories are close enough to the dots in the (cloud of) individuals and how relevant it is."*

Interactions. All three experts found the brushing and linking to be highly useful, though the hexbin plots were a little confusing for E1 and E2, who took a highlighted hexbin in one view to indicate that all the points in that bin were linked to the cluster selected in another view. E1 suggested providing *"a measurement of how much the correlation or lack of correlation is."* E3 initially found the Boolean operations to be less intuitive, but after asking for and seeing examples of how they were used, deemed the operations to be highly useful. Finally, E3 suggested the addition of numerical filters using which they could identify maintenance costs higher than a certain threshold, while E2 suggested filtering out data associated with commonly-occurring tags to help examine less-common problems.

Overall, the experts found the datatype-based separation less intuitive but considered the coordinated views and Boolean operations across the views to be of value. They recommended more tangible ways of grounding the data in the domain familiar to them by using locations of machines in buildings, locations of components in machines, and filtering by cost, dates, and keywords.

8 Discussion

The use case scenario and the expert review illustrate the importance of interactive visual analysis in the maintenance workflow. For instance, the overview visualizations were seen to provide useful groupings for analysts to explore and interpret using their domain knowledge. Additionally, the expert review illustrates the usefulness of interaction and filtering in helping interpret unfamiliar visual abstractions, and highlights a need to ground the representations in a way that is familiar to the domain experts.

In the use case scenario, we saw how the overview visualizations can help identify common patterns across the dataset (e.g., the *"hot/cold"* cluster) and help small but closely related clusters stand out (e.g., maintenance of equipment involving fans). The ccMCA views (Fig. 7) allowed the user to not only verify common traits—such as the presence of unreliable fans, or replacements ordered by a small subset of supervisors—across a problem group but also identify which equipment (in the case of fans) and supervisors (in the case of replacements) had the common traits.

The expert review highlighted both the advantages and disadvantages of our approach. When the domain experts were demonstrated scenarios, such as that described in Sect. 6, they were convinced and impressed by the capability of the prototype. Their validation of the workflow used to create the projected views and characterization views (Fig. 1) also verified that our approach was well-motivated. On the other hand, the experts found the data separation and visualization too abstract to pick up in a single session. They preferred a more tangible means of viewing the data, based on the location of the machines, locations of the components in the machines, and based on cost. However, representing high-dimensional data based on only one or two characteristics may not reveal important insights. In addition, one of the main advantages of our approach—its generalizability to other domains—will be lost by grounding it too much in one domain. However, there may be a middle ground wherein the user is able to add an additional "custom" view based on familiar data characteristics. We will explore this in future iterations.

In spite of the difficulty the experts faced with the abstract representations, they found the coordination or linking across views to be a useful feature that helped them understand the data better. As with the representations, they did express a preference for more tangible filters (e.g., based on specific cost ranges). However, at least one expert (E3) had started to appreciate the sophisticated filtering possible through the coordinated views and Boolean operations. The expert feedback suggested that some of what they found difficult about the interface was more due to the short duration of the sessions rather than the data abstractions themselves. A longitudinal study—though unrealistic at this time with restrictions on data sharing and the current constraint of remote sessions—would help address some of the familiarity issues that the domain experts currently face.

9 Conclusion

In this paper, we present a design that couples machine learning with interactive visualization for analyzing large, heterogeneous, multidimensional maintenance log data. A key approach is to separate numerical, categorical, and text dimensions of the data, and use lower-dimensional, clustered views that reveal groups in the dataset by each dimension type. We apply existing techniques such as ccPCA and word embeddings with frequency plots to characterize the dataset based on its numerical and text dimensions. Notably, a unique capability is provided with our new contrastive learning method, ccMCA, to characterize a dataset with its categorical dimensions. We present these approaches of clustering and characterization in the form of a dashboard with linked views, and illustrate its utility through a use-case scenario and an expert review. These scenarios allow us to highlight the use of ccMCA in identifying categorical dimensions and their values that contribute to a cluster. The expert review highlighted the usefulness of linked views to characterize clusters across different dimension types. We also identify the need for more grounded, domain-specific representations of data to scaffold the experts' understanding of the system.

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.

Acknowledgement

We are grateful to the domain experts who volunteered with our hour-long study during these challenging times. We also thank the anonymous reviewers for their suggestions that helped improve this paper. This research is sponsored in part by the financial assistance award 70NANB20H197 from the U.S. Department of Commerce, National Institute of Standards and Technology.

References

[1] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *Proc. HICSS*, pages 1–10. IEEE, 2011.

[2] A. Arleo, C. Tsigkanos, C. Jia, R. A. Leite, I. Murturi, et al. Sabrina: Modeling and visualization of financial data over time with incremental domain knowledge. In *Proc. VIS*, pages 51–55. IEEE, 2019.

[3] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, and G. Falkman. Interactive clustering: A comprehensive review. *ACM Computing Surveys*, 53(1):1–39, 2020.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D^3: Data-driven documents. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[5] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proc. BELIV*, pages 1–8, 2014.

[6] B. Broeksema, A. C. Telea, and T. Baudel. Visual analysis of multi-dimensional categorical data sets. *Computer Graphics Forum*, 32(8):158–169, 2013.

[7] M. P. Brundage, K. Morris, T. Sexton, S. Moccozet, and M. Hoffman. Developing maintenance key performance indicators from maintenance work order data. In *Proc. MSEC*. ASME, 2018.

[8] M. P. Brundage, T. Sexton, M. Hodkiewicz, K. C. Morris, J. Arinez, et al. Where do we start? guidance for technology implementation in maintenance management for manufacturing. *J. of Manufacturing Science and Engineering*, 141(9):091005, 2019.

[9] M. Cammarano, X. Dong, B. Chan, J. Klingner, J. Talbot, et al. Visualization of heterogeneous data. *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1200–1207, 2007.

[10] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield. Scatterplot matrix techniques for large N. *J. of the American Statistical Association*, 82(398):424–436, 1987.

[11] T. P. Carvalho, F. A. Soares, R. Vita, R. d. P. Francisco, J. P. Basto, and S. G. Alcalá. A systematic literature review of machine learning

methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024, 2019.

[12] G. Y.-Y. Chan, P. Xu, Z. Dai, and L. Ren. ViBR: Visualizing bipartite relations at scale with the minimum description length principle. *IEEE Trans. on Visualization and Computer Graphics*, 25(1):321–330, 2019.

[13] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Peppler, N. Elmqvist, and K. Ramani. VizScribe: A visual analytics approach to understand designer behavior. *International J. of Human-Computer Studies*, 100:66–80, 2017.

[14] M. Choi, C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proc. CHI*, pages 1–12, 2019.

[15] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *Proc. AMS Conf. on Math Challenges of the 21st Century*, 2000.

[16] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, pages 226–231, 1996.

[17] R. Faust, D. Glickenstein, and C. Scheidegger. DimReader: Axis lines that explain non-linear projections. *IEEE Trans. on Visualization and Computer Graphics*, 25(1):481–490, 2018.

[18] C. Felix, A. Dasgupta, and E. Bertini. The exploratory labeling assistant: Mixed-initiative label curation with large document collections. In *Proc. UIST*, pages 153–164, 2018.

[19] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan. ChronoViz: a system for supporting navigation of time-coded data. In *Proc. ACM Extended Abstracts on CHI*, pages 299–304, 2011.

[20] T. Fujiwara, O.-H. Kwon, and K.-L. Ma. Supporting analysis of dimensionality reduction results with contrastive learning. *IEEE Trans. on Visualization and Computer Graphics*, 26(1):45–55, 2020.

[21] T. Fujiwara and T.-P. Liu. Contrastive multiple correspondence analysis (cMCA): Using contrastive learning to identify latent subgroups in political parties. *arXiv:2007.04540*, 2020.

[22] T. Fujiwara, Shilpika, N. Sakamoto, J. Nonaka, K. Yamamoto, and K.-L. Ma. A visual analytics framework for reviewing multivariate time-series data with dimensionality reduction. *IEEE Trans. on Visualization and Computer Graphics*, 27(2):1601–1611, 2021.

[23] C. Guo and F. Berkhahn. Entity embeddings of categorical variables. *arXiv:1604.06737*, 2016.

[24] H. Guo, S. Di, R. Gupta, T. Peterka, and F. Cappello. La VALSE: Scalable log visualization for fault characterization in supercomputers. In *Proc. EGPGV*, pages 91–100. Eurographics, 2018.

[25] F. Heimerl and M. Gleicher. Interactive analysis of word vector embeddings. *Computer Graphics Forum*, 37(3):253–265, 2018.

[26] M. Hodkiewicz and M. T.-W. Ho. Cleaning historical maintenance work order data for reliability analysis. *J. of Quality in Maintenance Engineering*, 22(2):146–163, 2016.

[27] X. Jin, B. A. Weiss, D. Siegel, and J. Lee. Present status and future growth of advanced maintenance technology and strategy in us manufacturing. *International J. of Prognostics and Health Management*, 7(Special Issue on Smart Manufacturing PHM), 2016.

[28] P. Joia, F. Petronetto, and L. G. Nonato. Uncovering representative groups in multidimensional projections. *Computer Graphics Forum*, 34(3):281–290, 2015.

[29] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, et al. Clustervision: Visual supervision of unsupervised clustering. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):142–151, 2017.

[30] B. Le Roux and H. Rouanet. *Quantitative Applications in the Social Sciences: Multiple Correspondence Analysis*. SAGE Publications, 2010.

[31] P. Lindstrom, D. Koller, W. Ribarsky, L. F. Hodges, N. Faust, and G. A. Turner. Real-time, continuous level of detail rendering of height fields. In *Proc. SIGGRAPH*, pages 109–118, 1996.

[32] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Trans. on Visualization and Computer Graphics*, 23(3):1249–1268, 2017.

[33] E. Loper and S. Bird. NLTK: The natural language toolkit. In *Proc. ETMTNLP*, pages 63–70. ACL, 2002.

[34] S. Lukens, M. Naik, K. Saetia, and X. Hu. Best practices framework for improving maintenance data quality to enable asset performance analytics. *Annual Conference of the PHM Society*, 11(1), 2019.

[35] E. Mahfoud, K. Wegba, Y. Li, H. Han, and A. Lu. Immersive visualization for abnormal detection in heterogeneous data for on-site decision making. In *Proc. HICSS*, 2018.

[36] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[37] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist. ConceptVector: Text visual analytics via interactive lexicon building using word embedding. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):361–370, 2017.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. Scikit-learn: Machine learning in python. *J. of Machine Learning Research*, 12:2825–2830, 2011.

[39] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of LIWC2007. LIWC Manual, `https://www.liwc.net/LIWC2007LanguageManual.pdf`. Accessed: 2020-4-30.

[40] K. Reda, A. Febretti, A. Knoll, J. Aurisano, J. Leigh, et al. Visualizing large, heterogeneous data in hybrid-reality environments. *IEEE Computer Graphics and Applications*, 33(4):38–48, 2013.

[41] R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proc. LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[42] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, et al. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):241–250, 2017.

[43] T. Sexton, M. P. Brundage, M. Hoffman, and K. C. Morris. Hybrid datafication of maintenance logs from ai-assisted human tags. In *Proc. IEEE Big Data*, pages 1769–1777, 2017.

[44] T. Sexton, M. Hodkiewicz, and M. P. Brundage. Categorization errors for data entry in maintenance work-orders. *Proc. PHM*, 11(1), 2019.

[45] T. B. Sexton and M. P. Brundage. Nestor: A tool for natural language annotation of short texts. *J. of Research of National Institute of Standards and Technology*, 124:124029, 2019.

[46] Shilpika, B. Lusch, M. Emani, V. Vishwanath, M. E. Papka, and K.-L. Ma. MELA: A visual analytics tool for studying multifidelity hpc system logs. In *Proc. DAAC*, pages 13–18. IEEE, 2019.

[47] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan. Big data clustering: A review. In *Proc. ICCSA*, pages 707–720, 2014.

[48] R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proc. AAAI*, pages 4444–4451, 2017.

[49] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):629–638, 2016.

[50] P. J. Stone, D. C. Dunphy, and M. S. Smith. *The general inquirer: A computer approach to content analysis.* MIT press, 1966.

[51] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820, 2014.

[52] G. K. Tam, V. Kothari, and M. Chen. An analysis of machine-and human-analytics in classification. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):71–80, 2017.

[53] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proc. ACL*, pages 384–394, 2010.

[54] L. van der Maaten, E. Postma, and J. van den Herik. Dimensionality reduction: A comparative review. *J. of Machine Learning Research*, 10:66–71, 2009.

[55] J. Xia, W. Chen, Y. Hou, W. Hu, X. Huang, and D. S. Ebertk. DimScanner: A relation-based visual exploration approach towards data dimension inspection. In *Proc. VAST*, pages 81–90. IEEE, 2016.

[56] P. Xu, H. Mei, L. Ren, and W. Chen. ViDX: Visual diagnostics of assembly line performance in smart factories. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):291–300, 2016.

[57] M. X. Zhou and S. K. Feiner. Data characterization for automatically visualizing heterogeneous information. In *Proc. InfoVis*, pages 13–20, 1996.

Proceedings of the ASME 2020
International Mechanical Engineering Congress and Exposition
IMECE2020
November 16-19, 2020, Portland, OR, USA

IMECE2020-24037

DIGITAL TWIN FOR SMART MANUFACTURING: THE PRACTITIONER'S PERSPECTIVE

Maja Bärring[1], Björn Johansson
Chalmers University of Technology
Industrial and Materials Science
Hörsalsvägen 7A, 41296 Gothenburg, Sweden

Guodong Shao
National Institute of Standards and Technology
100 Bureau Drive, MS 8260
Gaithersburg, MD 20899, USA

ABSTRACT

The manufacturing sector is experiencing a technological paradigm shift, where new information technology (IT) concepts can help digitize product design, production systems, and manufacturing processes. One of such concepts is Digital Twin and researchers have made some advancement on both its conceptual development and technological implementations. However, in practice, there are many different definitions of the digital-twin concept. These different definitions have created a lot of confusion for practitioners, especially small- and medium-sized enterprises (SMEs). Therefore, the adoption and implementation of the digital-twin concept in manufacturing have been difficult and slow. In this paper, we report our findings from a survey of companies (both large and small) regarding their understanding and acceptance of the digital-twin concept. Five supply-chain companies from discrete manufacturing and one trade organization representing suppliers in the automotive business were interviewed. Their operations have been studied to understand their current digital maturity levels and articulate their needs for digital solutions to stay competitive. This paper presents the results of the research including the viewpoints of these companies in terms of opportunities and challenges for implementing digital twins.

Keywords: Digital Twin, Digitalization, Smart Manufacturing, Supply Chain

1. INTRODUCTION

The main focus and simultaneous challenge of manufacturing companies today is to produce products efficiently and deliver products on time. The major high-level challenges include (1) the dynamic requests for customized products and order changes and (2) the unpredictable status changes of processing equipment. Both challenges complicate the decision-making processes needed for manufacturing organizations to achieve their objectives [1].

A survey performed by the Supply Chain Management (SCM) World [2] provides a more detailed look at those challenges. Details include (1) shortening the response time to unforeseen events, (2) meeting the delivery time, (3) reducing the cycle time for introducing new products to market, and (4) enabling a flexible product mix. The survey clearly indicated that the current, rapid, IT-technology development has the opportunity to provide practitioners with more powerful and less costly technological tools to address these details.

By using these tools, practitioners can connect production systems digitally using the data collected from processes and humans in those systems. Such digital connection will provide opportunities to learn, to control, and to predict the system's future behavior. In this paper, we focus on one of the most promising of those technology tools, the digital twin [3]. Digital twin was listed as one of the Gartner Top 10 Strategic Technology Trends for 2019 [5]. The Gartner Top 10 Strategic Technology trends highlight changing or not yet widely recognized trends that will impact and transform industries by 2023 [6-7].

Currently, the digital-twin concept and its associated enabling technologies are not being implemented or embraced as rapidly and efficiently as expected [4], especially considering the hype and buzz about the concept. In our view, this situation is caused partly by the confusion about (1) what a digital twin actually is, (2) what it should include, and (3) where it should start to be implemented. Partly also by the need to collect, analyze, communicate, simulate, and integrate data in real-time. Most companies do not have the resources and expertise to address these needs. This is especially true for the small- and medium-sized enterprises (SMEs). They lack the definitions of the concept and associated challenges of digital twins. SMEs also need the information on enabling technologies and relevant standards, and systematic procedures for implementing digital twins.

[1] Contact author: maja.barring@chalmers.se

Barring, Maja; Shao, Guodong; Johansson, Bjorn. "Digital Twin for Smart Manufacturing: The Practitioner's Perspective." Presented at 2020 ASME International Mechanical Engineering Congress and Exposition (IMECE 2020). November 16, 2020 - November 19, 2020.

At the 2019 Winter Simulation Conference, a panel of experts from government, industry, and academia gathered to discuss this issue of implementing digital twins from a simulation perspective [4]. The panel summarized its discussion in two parts. First, there are needs of the digital-twin concept that go beyond the capabilities of traditional, simulation methods and tools. Examples of those needs include the real-time model execution and real-time two-way communication. Second, and consequently, each digital twin is unique in both the resulting model itself and how it must be used in a real manufacturing environment. That is, a digital twin is context dependent. Because of its summary, the panel concluded that the perspectives of manufacturers and other practitioners are especially important for the research community to address those two needs.

Practitioners are not the only ones enamored by the digital-twin concept and its enabling technologies. Both have been discussed greatly by researchers, academics, technology and solution providers, and Standard Development Organizations (SDOs). Similarly, most of those discussion, and a plethora of publications, focus on (1) what a digital twin is and (2) what it should provide. Unfortunately, the perspective of practitioners, are usually not included in those technological discussion and research publications. Real manufacturing use cases are rarely parts of those discussions and publications.

Consequently, answers to those two questions vary considerably across the communities. Interestingly, there is one place where all interested parties can convene, the International Organization for Standardization (ISO). ISO is developing a standard ISO 23247, "Digital Twin Framework for Manufacturing." This standard will provide a generic guideline and a reference architecture for case-specific, digital twin implementations [8]. The findings of the study in this paper may be valuable for the development of that standard. In return, the completed standard will guide the implementations of digital twins in manufacturing.

In this paper, the digital-twin concept, from manufacturers' perspective, is researched and investigated based on specific use cases from the participating organizations. The paper summarizes our multi-phased investigations into how those organizations view the digital-twin concept, its definitions, its applicability, and its implementations. The investigation was based on interviews of key persons in those organizations and visits to their plants to understand their current status on digital maturity level. In the first phase, we focused on their willingness and readiness to implement digital twins. In the second phase, we asked more detailed questions about their challenges with actually implementing digital twins. At the final stage, we focused on the interviewers' understanding of the technologies and standards needed to create and integrate digital twins into their operations, and if they have any procedures to follow for implementing digital twins. This paper summarizes the findings and also discuss how the adoption and implementation of the digital-twin concept can be expedited.

The remainder of this paper is organized as follows: Section 2 provides some background information about digital twins. Section 3 introduces the research methods for this study. Section 4 summarizes the findings regarding the current, digitalization level, and the interview results for each company. Section 5 discusses the benefits and the limitation of digital twin implementations, and Section 6 concludes the paper and discusses the future work.

2. BACKGROUND

2.1 Digital Twin Definitions

The term "digital twin" was first introduced in 2002 [9]. Dr. Grieves used it to explain how to create a "digital informational construct of a physical system as an entity on its own". That digital construct is a "digital twin" that represents all the embedded information in that physical system. Furthermore, this digital twin should be connected with the physical system throughout its entire life. Figure 1 shows a graphical representation of the digital-twin concept that includes the two spaces and the flows that connect them. According to [9], those flows are the essential part of a digital twin. Those flows involve two-way connections to exchange data and information between the two spaces. The information flow from the virtual space to the real space may be indirectly through the involvement of human decision makers.

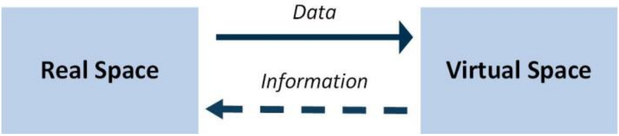


FIGURE 1: THE DIGITAL TWIN CONCEPT [9]

The real space may be a product, a process or a system. Depending on the context, associated digital twins should be able to fully describe the information embedded in either its potential or current physical instantiation to fulfil the objectives of the manufacturing problem. Therefore, a digital twin is a purpose-driven, virtual, complete or a partial representation of the product's, the process's, or the system's physical reality. For example, much of the information embedded in a physical product can be observed by physically inspecting it. Or, it should also be possible to predict that same information from the digital twin of it [9].

In practice then, a digital twin is context-dependent, purpose-driven, often time, an incomplete representation of any physical reality, i.e., a product, a process, or a system. Because of this, a digital twin implementation only requires the data and the models associated with the intended purpose [10]. Usually, that purpose is to optimize one or more performance metric(s). Achieving that purpose requires the real-world data needed for tracking the behavior and predict the performance of the specific element in the real space. The tracking and analyzing are done using digital twin models in the virtual space. A basic, enabling technology for connecting these two spaces is the Internet of Things (IoT).

Other than requiring a relatively large amount of IoT input data, a digital twin model also needs the computing power for (1) processing and analyzing the data and (2) executing the digital twin model that uses the data to track and predict the chosen performance metrics. To gain the full potential of a digital twin, it should use the results of the model to make decisions and to take the actions autonomously. These decisions and actions will help improve the performance of the physical counterpart. For example, inspection data can be used with a digital twin model to optimize product metrics. In another example, operational data can be used with digital twin models to lay the groundwork for customized, mass production techniques [7].

Another way to describe a digital twin is by its capabilities, including connectivity, visibility, granularity, and analyzability [4]. Connectivity is about the level of data exchange that exists between the digital twin and its physical counterpart. Visibility is about the ease with which a human being can understand the digital twin results. Granularity is about the level of detail in the digital twin. Analyzability is about the ability of the digital twin to support decision-making.

As noted, the digital twin terminology has been used since 2002. Long before that, however, there was an idea for digital representation of a production system. A representation with the ability to be updated with real-time information. Fisher [11] explained in 1986 that a virtual factory is an “electronic model” of the real system with ability to be updated. In the manufacturing domain, the idea of a virtual factory and digital twin do largely overlap.

2.2 Challenges with Implementing Digital Twins

Building a digital twin requires a homogenous perspective of the information that can persist across the organizational functional boarders. Implementing such a perspective, requires overcoming five, existing obstacles that are making it difficult for manufacturers to effectively implement a digital twin. First, the information associated with organizational functions (such as design, engineering, and manufacturing) is stored in silos. This significantly limits information sharing between functions. Second, there is a lack of detailed knowledge about the physical world. That knowledge is needed to understand and simulate the natural phenomena as it becomes virtually represented. Third, there is uncertainty about the possible states that a system can take and the time it will be in that state [9]. Fourth, it is difficult to include human involvement, which happens in most manufacturing and logistics decisions, in the digital twin. And, fifth, the needed digital data to represent the actual physical system is largely unavailable in SMEs [4].

SMEs normally specialize in one specific market sector and its specific supply chains. Because of this, there is a risk an SME may limit its competitiveness and be more exposed to the market’s volatility. An alternative is to adapt a “plug and play” approach, where SMEs can be a part of various supply chains across multiple sectors. This allows them to take advantage of any available capacity and identify new business opportunities. To market their services to Original Equipment Manufacturers (OEM), SMEs must have the ability to describe their manufacturing capabilities effectively and efficiently [12]. Digital twins can provide that ability.

3. RESEARCH METHODS

This multi-case study involved visiting, and interviewing key personnel from, five companies in two supply chains: automotive and heavy vehicle. Because of this setup, the study focused on issues related to building digital twins for each company individually and on issues related to building each digital twin within the supply chains. The five companies are geographically spread and represent both OEMs and SMEs. A trade organization has also been interviewed to gain an extra perspective to this topic.

Figure 2 shows the research methods. Our approach to the study involved qualitative data collection from our observations of each company’s current operations and our interviews with the people running those operations. The agenda was the same for each visit. First, tour the production or logistics facility by following the product flow. Second, conduct a question-and-answer session between plant engineers, academic researchers, and technology providers to discuss the formal process flow. Third, conduct one-on-one, semi-structured interviews with the key personnel from each company to explain their current understanding and status of digital twin implementations.

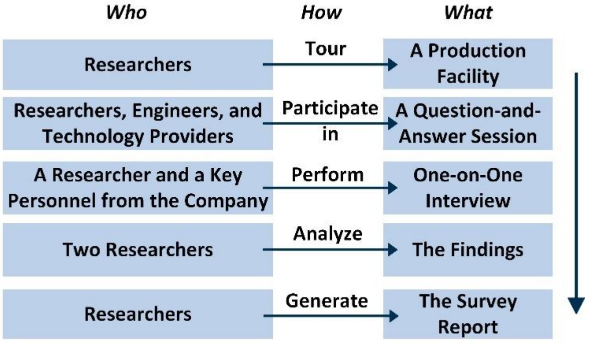


FIGURE 2: THE RESEARCH METHODS

The semi-structured interviews were planned for 30 min and the survey questionnaire mainly covered five topics (1) their definitions of a digital twin, (2) their prerequisites and needs for digital twins, (3) the key technologies for digital twin implementations, (4) which lifecycle problems are best suited digital twin solutions, (5) challenges involved and their future directions for building digital twins. The interviews were performed by one researcher interviewing one company representative.

The interviews were recorded and transcribed. The transcribed interviews together with notes taken during interviews were the basis for our analysis. All interviewees possessed great knowledge about their production processes and their organizational structures. For the analysis, the categories identified from literature as of importance to the digital twin concept were used for scoping the interviews. The transcribed interviewees' answers were analyzed by two researchers to categories them and identified the common denominators and differences in the answers.

4. THE PRACTITIONER'S PERSPECTIVE

This section first introduces the participating case companies and the two involved supply chains. The section describes each company including (1) the products it makes, (2) the processes it uses to make the products, (3) the way it currently communicates shared information, and, (4) the challenges and problems it faces as a single partner in a more complex supply chain. Then, the section summarizes the results of the interviews and, finally, explores those results from both perspectives: an individual one and a supply chain one.

4.1 Who are the Case Companies?

As noted, the case companies come from two different industries: automotive vehicles and heavy vehicles. The companies vary in size, ranging from 30 employees to thousands of employees. They vary in production-related capabilities, starting from mechanical workshop where one-off products can be manufactured; to the logistics, raw material, and components providers needed to assemble forklift trucks and part for the automotive industry; and, a highly automated plant with robot-welding.

The core competencies of the case companies also vary depending on their end-customers. For the automotive sector, logistics expertise and streamlined processes are keys. For the heavy-vehicle sector, the level of flexibility required to meet the end-customers' needs is key. The descriptions of the five case companies and their two supply chains are summarized in Table 1, where CC stands for case company and SC stands for supply chain. Industry indicates their sectors, and the size covers large (L) or medium (M) or small (S) enterprise. One additional organization is Company F, which is a trade organization representing the interest of hundred suppliers to the automotive industry. Many of Company F's members are SMEs and case companies involved in the project are member of Company G.

TABLE 1: SUMMARY OF PARTICIPATING CASE COMPANIES

CC	SC	Industry	Size
A	1	Heavy vehicles	L
B		Forging of steel	M
C		Mechanical workshop	S
D	2	Automotive	M
E		Customizing steel products	S
F		Trade organization for suppliers in the automotive industry	

- *Supply Chain 1*

Companies A, B, and C belong to the heavy-vehicle supply chain. These companies range in size from large (Company A) to medium (Company B) and small (Company C). Company A designs and manufactures forklift trucks for customers who handles shipping containers at harbors. Forks are the most customized parts because they must be designed based on their application and intended use. Parameters include weight to be handled, reach of forks, and area to be operated in. Since the demand on customization is high and is one of the company's competitive advantages, their suppliers of forks are required to provide a high flexibility for the kind of fork to be manufactured.

The supplier of the forks in this chain is Company B. Company B must purchase the required materials well in advance because the material has a long lead time. The production system is highly flexible with the capability to forge almost any kind of fork structure. The production system has a functional layout where the same functions are organized to be in the same area. The layout is also organized in a way to support the product flow through the production site.

Company C is a mechanical workshop that manufacturers fork components for Company B. The company can supply one-off products based on customers' needs. Company C also has a functional layout of the production system where the same functions (e.g., computer numerical control (CNC) machining) are organized in the same area.

Each company in this supply chain still has its own internal, information silos. The companies do have Enterprise Resource Planning (ERP) systems to manage their internal information. However, communicating and sharing downstream, order-related information among them is done semi-automated or by emails. Company A can send its ERP related order information directly and electronically to Company B who receives it as an email; and the electronic communication downstream to Company C is also handled indirectly by e-mail.

Currently, there are no direct, automated ways to communicate and share upstream information. That happens using phone, fax, or email. This competitive disadvantage is far outweighed by a significant, competitive advantage in this chain, its flexibility. The OEM is able to deliver the right product, at the right time, and based on customer specifications. This implies that the current, technologically limited, flow of information can effectively handle the dynamic, specification changes that customers make to their products. Changes that will require new product, new design, and new production information to flow through the supply chain.

- *Supply Chain 2*

Supply chain 2 comprises two companies: D and E. Company D mostly supplies to OEMs, with a large market share in the automotive industry. Company E supplies to Company D, but it also has a broader customer base that crosses industry sectors. In this supply chain, Company E supplies customized steel products that are created from the steel coils that are received from steel producers. Company E can (1) secure a delivery precision that the steel manufactures do not do, (2) customize the coils based on specification, and (3) supply material from numerous steel manufactures globally. Its production facility contains a warehouse area to store coils and several machines that make steel rolls and sheet metals.

As noted, Company D delivers manufactured parts to the automotive sector. These parts are used for supporting other components to cabs. Their production layout is based on the product flow and operations include pressing, welding, painting, and assembly. Since many operations have become more automated, the role of the operator has changed from operating the equipment to more supervising the process, handling the material flow, and ensuring the quality of the products.

The information exchanged in this supply chain is done using the Electronic Data Interchange (EDI) standard. EDI is a commonly used standard in automotive supply chains. It is enforced by OEMs. This in turn encourages at least, the top two tiers of the supply chain to use EDI to communicate and share information.

The two supply-chain strategies are similar. The OEM places an order based on a product prognosis that use historical data from their various customers. If that order includes customized products, there will be a long lead time for the customer. If the order includes only standardized products, then the first option is to take the products from stock. If no stock, then place orders to the first-tier suppliers using EDI. Both companies, D and E, are using Monitor as an ERP system. So, they can easily exchange order information.

Company D has a high degree of automation in all its production processes; Company E has implemented an automation solution for material handling. From the perspective of the typical, ISA 95, plant hierarchy, there is an information-processing gap between the ERP level at the top and the low levels close to the real production. In the ISA-95 view, that information processing is done in a Manufacturing Execution System (MES) level. In both companies, that level is missing. Therefore, production equipment and personnel are still not connected to provide data to support decision-making.

- *Trade organization for suppliers in the automotive industry*

Company F, a trade organization in the automotive industry, is also another entity in this study. Many of its member companies are SMEs; therefore, Company F can represent the views and needs of a large group of companies regarding digitalization. Because of their special position of representing the interest of hundreds of companies, trade organizations are actively involved in research programs, projects, and initiatives. Two big trends that they currently emphasize are the electrification of vehicles and digitalization related to products and production.

4.2 What are their Viewpoints on Digital Twins?

The interview results are summarized in Tables 2, 3, and 4. They are grouped based on various topics. Table 2 includes the answers from the case companies regarding the terms, enabling technologies, prerequisites, possible life-cycle stages, relevant standards, and procedures for building digital twins. Table 3 focuses on opportunities and challenges involved for developing and implementing digital twins. Table 4 is about a general, digitalization strategy, smart manufacturing technologies, and acceptance for new ideas on digitalization. Each table is organized by responses from the six case companies, i.e., each question is followed with the answers for all the companies.

Somewhat surprisingly, most companies were aware of the term “Digital Twin” before the interviews. The main reason was that they were introduced to the term in late 2018, when the companies first participated in a related research project (Digitala Stambanan, i.e., the National Digital Highway). Two interviewees reported that they were new to the concept.

Regarding what is a digital twin, the common view of the companies is that a digital twin is a digital and virtual representation of a factory or a product. Two synonyms for a digital twin - a “digital copy” and a “simulation” were also emphasized. Companies listed several technologies needed to develop and implement digital twins. Those technologies are (1) software applications to build digital-twin models, (2) sensors and actuators for collecting data and controlling physical elements, (3) 3D laser scanning for generating a virtual representation, and (4) computing power for handling large datasets.

Most interviewees agreed that standards could be helpful for implementing these technologies. Relevant standards would also give each company a common ground upon which it can (1) start its own implementation efforts and (2) share “best practices” with other companies attempting to do the same.

As for the potential, digital twin application life-cycle stages, production and maintenance are the most mentioned. This is not surprising since simulations are commonly used to analyze the performance impacts of different production flows prior to implementing one in real world. Additionally, simulations are often used for preventive maintenance by predicting when machines will fail. One interviewee expressed the view that there would be a digital twin potential in the design stage as well, i.e., the company may implement digital twins in all three stages. Interviewees said that their next steps would be a detailed research into the digital-twin concept and its enabling technologies. Companies would then use their research results to acquire the resources and technologies needed to build and implement digital twins. One interviewee mentioned that they already had ongoing work on digital twin development.

As to the potential benefits of implementing digital twins, the general opportunities that digitalization brings were mentioned, for example, the capability of representing and documenting the reality. Also, capabilities of performing predictive maintenance, simulating and testing different scenarios, and enabling changes to the product prior to production are listed. In addition, enabling different functions to work more closely, digitally, and removing barriers of geographical distances allows a holistic view for the enterprise or supply chain.

Challenges associated with implementing digital twins involve the right competencies; capable resources; cost and time; and also gaining interest and acceptance within the company, especially the support of management. Overall, interviewees emphasized more on the costs than the other challenges. Two interviewees said that because automotive suppliers have small profit margin, they would not quickly make new and larger investments on digital-twin development. There were several discussions regarding the possibility that other industries would adopt this concept earlier than automotive industry. One interviewee did not see any challenges and expressed: *"there are only opportunities."*

Lastly, general digitalization areas were investigated. One of the questions focuses on the case companies' general plan and strategy for digitalization. Many of the interviewees expressed that they do not have a strategy in place; however, they did try to follow and adopt the latest technologies. Some of them also mentioned that they, as a SME, can make decisions quick and easy (the hierarchy is not a hinder) and that they have Chief Executive Officers (CEOs) involved in the loop to keep them to stay relevant. The trade organization, Company F, expressed that their members typically have no strategy for making the changes needed for digitalization.

To get management to accept a change is normally by (1) showing examples and demonstrating the opportunity, (2) explaining Return on Investment (ROI), (3) stimulating a genuine interest for the area, and (4) involving management in the entire process. Customers can easily get management to accept a change by demanding a change in order to preserve a continued, or to create a new, business partnership. A common demand customers make involving the use of a particular set of standards, in particular, quality standards. The last question focused on the two digital-twin required technologies that members are not using. First, most members have no smart sensors in their plants. They normally store any data collected in their plants locally on a server. Second, their only models are mostly CAD representations of a machine or equipment.

5. DISCUSSION

This Section is divided into five main areas (1) definition of a digital twin, (2) potential benefits of implementing a digital twin, (3) challenges involved with digital twin development, (4) future directions of digital twins in manufacturing, and (5) quality of the research method.

5.1 Definition of a Digital Twin

A digital twin has been defined as digital representation that is created in a virtual space as a "twin" that represents the embedded information of a physical-space element throughout its entire lifecycle. The information is communicated to its associated digital twin, where it is analyzed in the virtual space. The result of analysis, typically information about a decision, is then communicated back to its associated physical element. Therefore, different types of information are being exchanged continuously and bi-directionally between the virtual and real spaces.

For most of the interviewees, the concept "Digital Twin" was relatively new. Many acknowledged that before their involvement in the Digitala Stambanan project, they had not heard the term. A few of them only learned the term during this survey and believed not so many others in their organizations are familiar with it either. Those who did have prior knowledge about digital twins, intuitively believed that a digital twin somehow, involved a digital copy and a virtual representation of their physical systems. Additionally, they believed that a digital twin could help them simulate their systems and predict future events. Also, they suggested a digital twin could support them in understanding the performance impacts of the increasing granularity of their products, processes, and systems.

From the study, it is further proven that a digital twin is context-dependent and use case specific. Each company has its own domain and specialty where the manufacturing problems, although related, will not be exactly the same. The digital twin implementation, therefore, will be specific for its intended use, performance objectives, and modeled parameters. However, the general approach should be the same. Moreover, some of the common modules, enabling technologies, and standards may be reused. In addition, it is not necessary to develop a digital twin of an entire plant. Rather, a digital twin should focus on the specific physical-space problem, e.g., a material-flow problem. In this way, individual digital twins or their modules can be reused for similar, even larger, future projects.

5.2 Potential Benefits of a Digital Twin Implementation

The potential values of having a digital twin were acknowledged by all interviewees. They view digital twins as something that could help them develop and analyze their system better. Adopting the digital-twin concept can (1) revolutionize the way they are currently doing their business, (2) become a driver for a complete digitalization of their operations, and (3) contribute the advancement of manufacturing industry. It has these three benefits, because a digital twin provides a common environment with tools to support better decision-making, across the plant, the enterprise, and even the supply chain.

Currently, commercial digital twin solution providers such as SAP, Visual Components, and AnyLogic are helping their customers to implement case-specific digital twins [13-15]. In high level, the results promised by these solution providers align well with the digital twin implementation benefits expected by our case companies. However, developing digital twins using these specific solutions involves many challenges.

5.3 Challenges involved with Digital Twin Development

Organizational, information silos and lack of knowledge of the physical world are obstacles for digital-twin development. In the case companies, such information silos exist and there is also limited use of advanced technologies, IT, and process technologies. This implies, as noted above, that the types of digital data for representing the physical world are limited. Other limitations stressed by the interviewees include their knowledge about digital-twin concept and their abilities to sell the concept internally. Many of the case companies are SMEs without their own IT departments. Therefore, to proceed with this digital-twin concept, they would need to learn more about it and understand the resources (including IT resources) required to implement that concept. Finally, all agreed that the time and cost are the major stumbling blocks to a successful implementation of digital twins. Both must be offset by the potential for significant improvement in performance, quality, and ROI. And, those improvements must be demonstrated before acceptance from management will be achieved.

TABLE 2: THE DIGITAL TWIN CONCEPT

1. Have you heard about digital twin?	
A	Yes, at the beginning of the research project.
B	Yes, but not before the research project.
C	Yes, this topic could go beyond the manufacturing applications, e.g., it could be applied in the building sector.
D	No.
E	No, and there are not many people who are familiar with it within the company.
F	Yes.
2. In your opinion, what is a digital twin?	
A	Building the factory digitally, which creates a digital copy that describes the organization and can be used for simulations.
B	A computerized 3D copy of a product or a production system or equipment.
C	A digital copy of a machine for simulating tool wear for when it occurs and how it occurs.
D	A CAD model that can be shared.
E	To build your factory virtually that will allow you to plan well before implementing it.
F	It is a digital copy of something that exists in reality, it could be a production system or a product.
3. What do you think of the key technologies required for the digital twin implementations?	
A	Sensors and actuators for sensing the reality and copying it to the digital world.
B	3D laser scanning to create a virtual representation, software to handle data from equipment and an analysis tool.
C	Do not know.
D	CAD and tools for building a virtual representation of a production system.
E	Software to handle the digital twin.
F	Computing power, since it will be required to handle massive amount of data. Good quality of the data that will be input to the digital twin.
4. What do you think of the standards that could support the digital twin implementations?	
A	It can support, since it can give insights of what is already known and existing solutions. The alternative is that everyone works on their own solution. Standards can make it more aligned and increase integration between organizations.
B	Yes, it will give a starting point.
C	Standards are mainly used when it is required by a customer, e.g., for quality control.
D	Yes, but normally standards are requirements that we need to comply with based on customers' needs.
E	Not easy to answer and we will need to make a unique implementation of it. Standards are mainly used for quality today.
F	Yes, standards could be used much more.
5. If applied, which of the lifecycle stage (i.e., product design, production, maintenance, etc.) you care the most?	
A	Production, to simulate flows and specially to simulate it in advance. To utilize how we are already working today to come up with how we will do it in the future. We are currently working a lot on design and material flows. Solutions in one factory could be transferred to other factories as well.
B	Production, to simulate different scenarios; maintenance, to some extent, and design, to modify products.
C	Maintenance, including both production equipment and infrastructure such as ventilation.
D	Product, in pre-production stage.
E	Production and maintenance, an important aspect is the production flows and how that can become more efficient. Maintenance does not have a more strategic plan.
F	It has potential to be important for all three stages and it should probably be investigated more. But maintenance will probably be the most important stage and artificial intelligence (AI) is more important for production.
6. What would be the further steps to take for you to build a digital twin?	
A	This is already being investigated, both for products and production.
B	(1) Investigate resources, currently we do not have our own IT department and (2) investigate technology required.
C	Do not know, need to research more to understand it better.
D	Focus on the digital representations of our own products.
E	To get more insights of how it works, how you would proceed with the software needed, or if a service would be used. Also, getting more people within the company to involve in this.
F	Competence about it, it is very important, also that it can be afforded by the organization. Need to decide the timeframe to implement it.

Barring, Maja; Shao, Guodong; Johansson, Bjorn. "Digital Twin for Smart Manufacturing: The Practitioner's Perspective." Presented at 2020 ASME International Mechanical Engineering Congress and Exposition (IMECE 2020). November 16, 2020 - November 19, 2020.

TABLE 3: BENEFITS AND CHALLENGES INVOLVED WITH DIGITAL TWINS

7. What benefits do you see with digital twins – for your company?	
A	The digital world makes it possible to simulate, improve, and change in a positive manner. It is a way of documenting that can be utilized in new projects. In this way, the need of physical prototypes can be decreased.
B	For modifying a product in the pre-production phase.
C	Predictive maintenance, but there is much more.
D	It is always good to have a digital representation because it is precise and easily accessible.
E	Simulate and plan flows in order to see new opportunities. The better our flows become, the more effective we become. Also, it is useful for planning maintenance.
F	The benefit is that you can simulate and test things out without interrupting in the real production system. However, it will require a business model for it.
8. What benefits do you see with digital twins – in general for manufacturing industry?	
A	An opportunity is for different functions in an organization that might spread geographically can meet digitally. Also, to try out new ideas before implementing it in reality.
B	Simulate and test different scenarios of both product and production digitally before going live, it is useful for multiple industries and companies.
C	Simulate different scenarios that can improve your production settings. It will help save time and cost.
D	For product information and in the context of cost optimization, you can evaluate the design in advance and give suggestions to the customer easier and earlier. Better repetitive in processes and material efficiency.
E	You can see more and view it with a holistic view. You will receive a good overview.
F	Same as for question 7, but the automotive industry has small marginal, which means that they might be hesitant for new investments. Also, they are big and do not embrace new changes easily.
9. What challenges or obstacles do you see for digital twin implementations – for your company?	
A	Competence, we do not really have it internally. It may be more costly if we would use a service to develop a digital twin.
B	Resources and general IT related knowledge.
C	Costs involved.
D	No challenges or obstacles.
E	To sell the idea internally, but if it can be demonstrated how it can impact profit, it will be normally positive.
F	Competence, time, and money.
10. What challenges or obstacles do you see for digital twin implementations – in general for manufacturing industry?	
A	Same as for our company and organizations, you have to be willing to learn and gain new knowledge.
B	Cost and other practical matters.
C	Cost.
D	No challenges or obstacles, there are only opportunities.
E	Cost, the automotive industry has low marginal and can therefore not easily commit big investments and changes, and the same as for question 9, i.e., to sell the idea internally.
F	Same as question 9 and the technical prerequisites exist. However, the rollout of 5G is going slow in Europe.

TABLE 4: DIGITALIZATION STRATEGY AND SMART MANUFACTURING TECHNOLOGIES

11. What is your plan for digitalization?	
A	It is one of our four focus areas and customers have come to us requesting this.
B	No stated strategy since our IT is from external, but is now more controlled by the company consortium. A current plan is to introduce barcodes for tracking material and software updates.
C	No stated strategy, but need to follow the development. The ERP system is the single source of truth that is updated and CNC programs are updated when needed. Computer Aided Manufacturing (CAM) programs are still too costly for us.
D	No stated strategy, but try to follow the development in the subject. It is important for the quality assurance of the complete chain and it is also important for attracting the future work force.
E	We are privately owned and always need to be updated on what happens regarding digitalization to keep the competitive strength and stay flexible. Digitalization can support that.
F	There is none.
12. What is the best way to help your management accept a new idea?	
A	Difficult question, but we do have groups working on future development. We also have a drive to develop new solutions. We can get help from other companies that are more advanced on this.
B	Show examples and demonstrate how it works and also provide ROI calculation to identify the potential wins and benefits.
C	We are a small organization; this means that decisions do not need to go through many roles in the company. The CEO is engaged in finding new opportunities, which makes it easier for new ideas. But everything is connected to a possible payback and monetary benefit.
D	To have a genuine interest for something and help push the development is a part of every employee's responsibility. To involve externally in forums and fairs to learn the latest trends and developments.
E	The economy is mandating a lot and the best way is to demonstrate the monetary benefits. That is always positive.
F	Unfortunately, it is often that the customers' demand is a number one reason to make the move. Secondly, it is about going through the motivation phase, where you understand the need for change. For all the current big changes, a general problem is the low level of awareness. We are working actively to change this.
13. Do you have smart sensors, databases, and models?	
A	*Smart sensors:* not for production but used for products. *Databases:* in the cloud. *Models:* yes, a model is used, for example, when moving the production from one site to another.
B	*Smart sensors:* no. *Databases:* local server and external IT support. *Models:* no, most of the time, the customer provides the product information and there is only paper-based information of the product.
C	*Smart sensors:* no. *Databases:* local server and external IT support. *Models:* no, but drawings are digital.
D	*Smart sensors:* no. *Databases:* local server, not in the cloud. *Models:* yes, CAD models of more or less everything related to production and products.
E	*Smart sensors:* no. *Databases:* locally. *Models:* yes, a CAD model of the newest machine.
F	*Smart sensors:* no. *Databases:* if the member companies save data it is mostly done locally. *Models:* no.

To demonstrate and exemplify how a new technology can be used is very important. The research project developed a testbed environment where new digital technologies can demonstrate how it can be used in the manufacturing context. The research community in general needs to publish more relevant research to help expedite the acceptance of digital twins. SDOs also need to speed up their effort on relevant standards that can guide manufacturing industry in its use of digital twins. Of course, if their budget allows and management approves, they can hire the commercial solution providers to customize their digital twin implementation for them.

5.4 Future Directions of Digital Twins in Manufacturing

The interviewees acknowledged both a potential future use of digital twins and that standards can guide companies and provide a starting point. As one of the interviewees indicated "instead of inventing how to do it independently, it could better be spread and shared what the best practice is." Many of them stated that the future steps for them are to study more about digital twins and get more people engaged in it. They also stated, however, that the cost constraints of automotive industry may prevent the industry from taking the lead on this effort.

In our view, even though the case companies are from specific sectors and in specific markets, the findings are typical representations of the current state of the manufacturing sector in general. Therefore, future efforts will include (1) developing relevant standards to facilitate the implementation process, (2) advancing enabling technologies to make them cheaper and better, and (3) educating decision makers and implementers to promote the acceptance of the digital-twin concept.

5.5 Quality of the Research Methods

A qualitative approach was used in this study, including observations and interviews. The goal was to investigate how much practitioners and potential users understand the digital-twin concept and its potential uses in manufacturing. Even though the number of interviewees is relatively small, they come from typical industries and their companies range in size from large to medium and small. Therefore, their views are representative, generalizable, and relevant to all manufacturing industries globally. In the future, we plan to perform more studies with more data, more practitioners, and more industry sectors.

6. CONCLUSION

This paper reports the results of a study that investigated the practitioner's perspective on digital twins. The practitioner's perspective has not been previously investigated or documented. Nevertheless, it is crucial for (1) the evaluation of the current state of the digital twin implementations and (2) the acceptance of the concept by the manufacturing sector. A digital twin was described as a digital copy of products, processes, or systems that can be used for understanding and predicting future events. There are potential benefits of digital twins for gaining better understanding of product designs, production operations, and system performance to support decision-making. But challenges involve the availability of resources and cost involved for getting acceptance by the management within the organization for building a digital twin. The study also revealed that different manufacturers have different needs, therefore, their digital twin implementations will be use-case dependent and purpose driven. While the execution of the study started with the participating case companies in specific sectors and specific markets, the findings and implications are more broadly relevant and valuable to all industries. The result of the study can serve as useful inputs for subsequent standards and technology development.

To help speed up the acceptance and adoption of the digital-twin concept by manufactures, global research communities should continue developing definitions, frameworks, and case studies. SDOs should also make efforts on developing guidelines and procedures to help industry, especially SMEs to start the implementation process.

ACKNOWLEDGEMENTS

The work presented in this paper has been supported by the NIST Model-Based Enterprise (MBE) program and Area of Advance Production at Chalmers University of Technology as well as the research project “Digitala Stambanan” funded by the Swedish Agency for Innovation Systems (VINNOVA). The involvement from case companies answering the questions related to digital twin was crucial for this study to happen and the authors would like to acknowledge their support and involvement.

DISCLAIMER

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. No approval or endorsement of any commercial product by NIST is intended or implied.

REFERENCES

[1] Shao, Guodong and Kibira, Deogratias. “Digital Manufacturing: Requirements and Challenges for Implementing Digital Surrogates.” *Proceedings of the 2018 Winter Simulation Conference*. pp. 1226 -1237. Gothenburg, Sweden, December 9-12, 2018. DOI 10.1109/WSC.2018.8632242.

[2] Manenti, Pierfrancesco. “Becoming a Smarter Manufacturer: How the Internet of Things Will Change the World”, SCM World, Boston, MA. 2015. http://www.scmworld.com/wp-content/uploads/2016/11/Becoming-a-Smarter-Manufacturer.pdf.

[3] Bolton, David. “What are Digital Twins and Why Will They Be Integral to the Internet of Things?” Applause. 2016. https://www.applause.com/blog/digital-twins-iot-faq/.

[4] Shao, Guodong, Jain, Sanjay, Laroque, Christoph, Lee, Loo Hay, Lendermann, Peter, and Rose, Oliver. “Digital Twin for Smart Manufacturing: The Simulation Aspect.” *Proceedings of the 2019Winter Simulation Conference*. pp. 2085-2098. National Harbor, MD, December 8-11, 2019. DOI 10.1109/WSC40007.2019.9004659.

[5] Costello, Katie and Omale, Gloria. ”Gartner Survey Reveals Digital Twins Are Entering Mainstream Use.” Gartner. 2019. https://www.gartner.com/en/newsroom/press-releases/2019-02-20-gartner-survey-reveals-digital-twins-are-entering-mai.

[6] Panetta, Kasey. “Gartner Top 10 Strategic Technology Trends for 2019.” 2018. https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019/.

[7] Siemens. “Digitalization in industry: Twins with potential.” 2020. https://new.siemens.com/global/en/company/stories/industry/the-digital-twin.html.

[8] ISO. “ISO (DIS) 23247-1: Automation Systems and Integration - Digital Twin Framework for Manufacturing - Part 1: Overview and general principles”. ISO/TC 184/SC4/WG15. 2020.

[9] Grieves, Michael and Vickers, John. “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems.” *Transdisciplinary Perspectives on Complex Systems*. Springer, Cham (2016): pp. 85-113. DOI 10.1007/978-3-319-38756-7_4.

[10] Shao, Guodong and Helu, Moneer. “Framework for a Digital Twin in Manufacturing: Scope and Requirements.” *Manufacturing Letters* Vol. 24 (2020): pp. 105-107. DOI 10.1016/j.mfglet.2020.04.004.

[11] Fisher, Edward L. “An AI-based Methodology for Factory Design.” *AI Magazine* Vol. 7 No. 4 (1986): pp. 72-85. DOI 10.1609/aimag.v7i4.560.

[12] Helu, Moneer, Sobel, William, Nelaturi, Saigopal, Waddell, Russell, and Hibbard, Scott. “Industry Review of Distributed Production in Discrete Manufacturing.” *ASME Journal of Manufacturing Science and Engineering* MANU-19-1674 (2020): pp. 1-26. DOI: 10.1115/1.4046988.

[13] SAP. 2020. https://www.sap.com/products/digital-supply-chain/digital-twin.html.

[14] Visual Components. “Digital twins and virtual commissioning in Industry 4.0.” July 2019. https://www.visualcomponents.com/insights/articles/digital-twins-and-virtual-commissioning-in-industry-4-0/.

[15] Wilkinson, Gavin. “White paper: How create a digital twin.” AnyLogic. 2018. https://www.anylogic.com/blog/white-paper-how-to-create-a-digital-twin/.

Solar Cell Performance Measurements Under Artificial Lighting Sources

Behrang Hamadani
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, USA
Behrang.hamadani@nist.gov

***Abstract*—In recent years, there has been a growing interest in measurements of photovoltaic solar cells under ambient artificial lighting such as light emitting diode (LED) or fluorescent light sources. Certain classes of solar cells are considered very good candidates for energy harvesting from mostly visible ambient lighting for the purpose of powering internet-of-things devices. However, measurements of the irradiance of these light sources, a key requirement for characterization of solar cells, has been challenging because there are currently no reference solar cells offered by any metrology laboratory for low light artificial measurements. The current approach of using illuminance meters for measuring the irradiance can result in unacceptable discrepancies between different labs. In this work, we take the first steps in demonstrating that a reference solar cell can indeed be calibrated under a well-defined low-light spectrum and can be used to perform current vs. voltage measurements on any test device under any arbitrary low light spectrum yielding consistent results. This work also highlights the pitfall of using lux meters for measuring light intensity and instead advocates for use of an effective irradiance ratio.**

***Keywords*—Ambient light, indoor light, current vs voltage measurements, energy harvesting, irradiance, photovoltaics**

I. Introduction

With the growing interest in measurements and characterization of solar cells under artificial or natural low light conditions, better testing methods need to be established to minimize inter-lab discrepancies [1]–[3]. Some classes of solar cells, particularly the inorganic III-V materials such as GaInP and GaAs devices [4], [5], and a variety of organic or hybrid photovoltaic materials [6]–[10] are well-suited for energy harvesting under these lighting conditions. However, the lack of a widely-accepted standard reporting condition (SRC) for indoor PV measurements or lack of calibrated reference cells for low light measurements has forced many researchers to use illuminance meters for measurements of the irradiance and report values in lux (lx = lm/m^2) [6], [11]–[13]. Illuminance measurements in units of lx can lead to widely different outcomes between different laboratories when measuring the same device. Two different light sources with identical illuminance at the measurement plane can have substantially different irradiance output on the solar cell, resulting in different short circuit currents (I_{sc}) and other *I-V* curve parameters.

This work proposes a reference-cell-based method for measuring and characterizing solar cells under various indoor lighting conditions. This method requires selection and use of a reference irradiance spectrum with an absolute scale much like the SRC defined by the two air mass (AM) 1.5 spectra (global and direct). Since lux-based measurements have already become commonplace in the low-light *I-V* measurement community, one can still maintain a connection to them by designing a reference spectrum such that the total illuminance equals 1000 lx. The proposed approach significantly streamlines the process of measuring the *I-V* curves under any light irradiance and will allow for the precise computation of the power conversion efficiency (PCE), a determination that is difficult to achieve with low uncertainty with lux-based measurements.

II. Experimental Details

All measurements were performed inside a dark box with two different white LED light sources projected down onto a stage where both the reference and the test solar cells are placed side by side and under the complete illumination of the LED spotlight. The LEDs are fan-cooled and operated in DC mode with a computer-controlled LED driver. The setup is such that the irradiance level is first set based on the photocurrent measurements from the reference cell and the simultaneous effective irradiance computation. Then an electronic switch opens up the source meter to the test cell so that an *I-V* curve sweep can be performed on it, usually from the I_{sc} to the open circuit voltage, V_{oc}.

The light source spectra were measured using an FEL-calibrated spectroradiometer [14] having an uncertainty of about 0.5 % in the spectral range of interest. The differential spectral responsivity (DSR) method was used to calibrate a silicon reference solar cell under the two reference spectra devised for this work [15], [16]. The linearity of the reference cell under these low irradiance conditions was also investigated [17] and it was determined that this particular reference cell is sufficiently linear (lower than 1 % variation) under irradiance levels probed for this work.

III. Methodology

In order to calibrate a reference solar cell under an artificial light source, a reference spectrum with an absolute irradiance scale needs to be constructed. We constructed two reference spectra shown in Fig. 1 such that one has a correlated color temperature (CCT) [18] of 3000 K and the other has a CCT of 6000 K. Each spectrum has a total illuminance of 1000 lx (lm/m^2), conditions that are considered fairly typical of bright indoor lighting with a warm white and a cool white chromaticity. In order to construct the reference spectra, first a relative spectral irradiance curve $\hat{E}$ was selected based on measurements of a white LED source and then this curve was

Hamadani, Behrang. "Solar Cell Performance Measurements Under Artificial Lighting Sources." Presented at 47th IEEE Photovoltaic Specialists Conference (PVSC 47). June 15, 2020 - August 21, 2020.

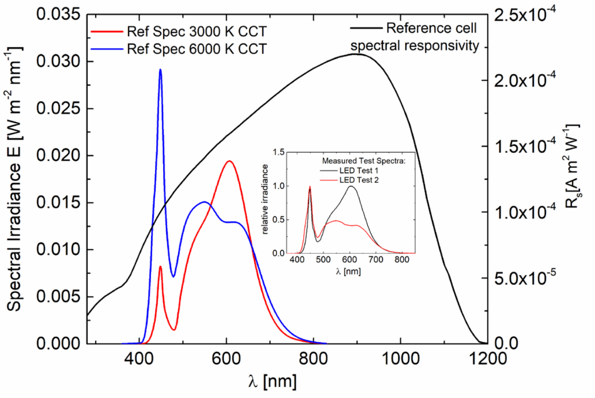


Fig. 1. The two reference spectra constructed for these measurements (left axis). The irradiance spectral responsivity of the reference cell (right axis). Inset shows the spectra of two test lights used for these measurements.

manually modified to have the intended color temperature with a CCT computational worksheet [18]. Then, $\hat{E}$ was multiplied by scalar α so that an absolute spectral irradiance curve $E_r(\lambda) = \alpha\hat{E}(\lambda)$ was computed in units of W.m^{-2}.nm^{-1}:

$$E_v = K_m \int_{360\text{ nm}}^{830\text{ nm}} \alpha\hat{E}(\lambda)V(\lambda)d\lambda\,, \tag{1}$$

where E_v is the illuminance set to 1000 lux, K_m is the spectral luminous efficacy for monochromatic radiation at 555 nm with a value equal to 683 lm/W, and $V(\lambda)$ is the normalized spectral luminous efficiency function, here chosen to be the photopic spectral luminous efficiency function with applications in high light levels such as bright indoor conditions or daylight [19].

In addition to establishing these reference indoor spectra, the spectra of test lights (LED, fluorescent or others) need to be measured, although not in an absolute way. The inset of Fig. 1 shows two test spectra corresponding to two different LEDs that were used as the actual illumination source for the measurements presented here. Within the mathematical framework described below, the spectral profile of the test light does not have to match the reference spectrum. Therefore, a different type of indoor light source such as a fluorescent lamp can be used for the *I-V* performance measurements.

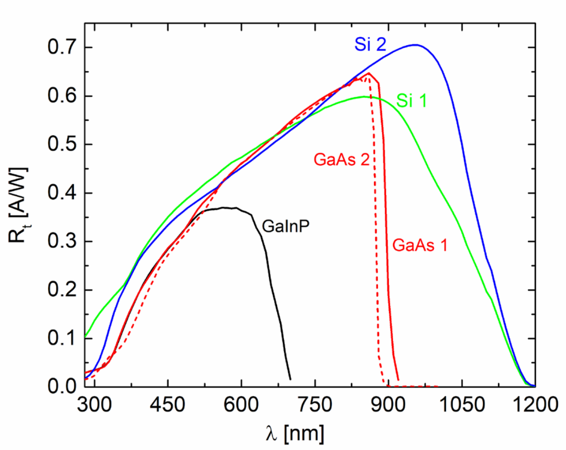


Fig. 2. The spectral responsivity curves of the solar cells tested.

The chosen reference cell is a silicon reference cell (nominal area 2 cm × 2 cm) at 25 °C and the DSR method was used to measure its irradiance spectral responsivity, $R_{r,irr}(\lambda)$. Then, the resulting short circuit current, $I_{r,r}$, under the reference spectrum at 25 °C is calculated using [15]:

$$I_{r,r} = \int_{\lambda\,\min}^{\lambda\,\max} E_r(\lambda)R_{r,\text{irr}}(\lambda)d\lambda\,, \tag{2}$$

where the subscript *r,r* stands for reference cell under reference condition. The right axis in Fig. 1 shows the irradiance spectral responsivity curve for the reference cell used in these measurements. The integral computation in (2) gives the following short circuit currents under the two reference spectra, $E_r(\lambda)$: $I_{r,r}^{CCT3k} = 444.98\ \mu\text{A}$, and $I_{r,r}^{CCT6k} = 517.70\ \mu\text{A}$.

A variety of solar cells were selected and characterized for this study. These devices included two types of silicon solar cells, labeled as “Si 1” and “Si 2”, one gallium indium phosphide (GaInP) solar cell and two types of gallium arsenide (GaAs) solar cells. Fig. 2 shows the spectral responsivity curves of the 5 test cells in units of A/W. All measurements are based on the DSR technique using a monochromator setup and dc light bias on each cell.

Given a reference spectral irradiance spectrum, a test spectral irradiance (the indoor simulator spectrum) and the spectral responsivities of both the reference and the test solar cells, a spectral mismatch correction parameter, *M*, can be calculated:

$$M = \frac{\int_{\lambda\,\min}^{\lambda\,\max} R_t(\lambda)E_t(\lambda)\mathrm{d}\lambda \int_{\lambda\,\min}^{\lambda\,\max} R_{r,irr}(\lambda)E_r(\lambda)\mathrm{d}\lambda}{\int_{\lambda\,\min}^{\lambda\,\max} R_t(\lambda)E_r(\lambda)\mathrm{d}\lambda \int_{\lambda\,\min}^{\lambda\,\max} R_{r,irr}(\lambda)E_t(\lambda)\mathrm{d}\lambda}\,, \tag{3}$$

where R_t is the spectral responsivity of the test cell, and E_t is the spectral irradiance of the indoor simulator source and can be unscaled. Recalling that the total irradiance incorporates an underlying spectrum to which the PV cells are responsive, one can define a unitless effective irradiance ratio, *F*, given by [20]:

$$F = M\frac{I_{r,t}}{I_{r,r}}\,, \tag{4}$$

where $I_{r,t}$ is the short circuit current of the reference cell under the testing condition, and $I_{r,r}$ is the short circuit current of the reference cell under the reference condition as obtained by (2). The measurement of *F* is readily achieved by monitoring the short circuit current of a calibrated reference cell mounted in the same incident plane of irradiance as the test cell.

With the goal of measuring the performance of these solar cells under the reference condition discussed above, (a) we placed both the reference and the test cells under the illumination source, i.e., indoor solar simulator, (b) calculate *M* for each pair, and (c) adjust the light levels while simultaneously reading $I_{r,t}$ and calculating *F* using (4) so that $F \to 1$ as best as possible. This may become an iterative process if the spectral shape of the test light changes with sourced current or neutral density filters. In this case, *M* has to be recalculated and *F* remeasured until it is sufficiently close to unity. Once satisfied, a 4-probe *I-V* curve is initiated on the test cell from 0 V to the open circuit voltage, V$_{oc}$, or in reverse if there are concerns about *I-V* curve hysteresis. The *I-V* curve

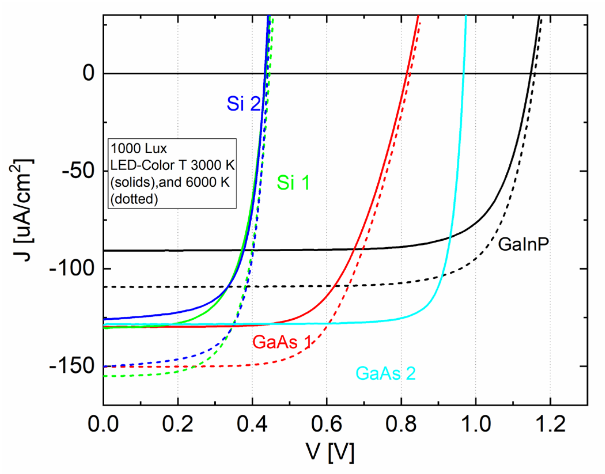


Fig. 3. J-V curves of five different solar cells under two constructed reference spectra, one with a CCT of 3000 K and the other with a CCT of 6000 K.

reported under the reference condition is given by:

$$I_{t,r}(V) = \frac{1}{M}\frac{I_{r,r}}{I_{r,t}} I_{t,t}(V) = \frac{S}{F} I_{t,t}(V)\,, \qquad (5)$$

where S is a uniformity factor defined as the ratio $I_{r,\text{ref cell pos}} / I_{r,\text{test cell pos}}$ for cases of nonuniform light sources.

Fig 3 shows the *J-V* curves for each of the 5 test solar cells described above under the established 3000 K CCT (solid curves) and the 6000 K CCT reference spectra (dotted curves), where J is the current density (=current divided by nominal cell area). For the GaAs2 cell, data was not available under the 6000 K CCT at the time of this submission. These measurements demonstrate that setting up the irradiance with a simple illuminance meter can lead to significant inconsistencies across different types of light sources and the shape of the spectrum plays a non-negligible role. For each of the test cells, it can be seen that the CCT 6000 K reference spectrum produces a higher magnitude J_{sc} and a slightly higher V_{oc} than the 3000 K CCT condition as a result of having a slightly larger total irradiance (≈ 3.7 W/m^2 vs 2.9 W/m^2, respectively).

When the test spectrum has a shape close to that of the reference spectrum, M ≈ 1 for most of the devices. However, when the test spectrum is cool white with a CCT of 6240 K but the intended reference condition has a CCT of 3000 K, then even silicon test cells show a mismatch parameter as high as 1.025. In general, the computations of all the measurement parameters show that if a spectral correction parameter is not computed for the measurement, then the potential errors are only minimized if either the indoor simulator's spectrum is chosen to be as close to the reference spectrum as possible, or the test and the reference cells are matched. If an inter-lab comparison is being performed when a lux meter is the only available instrument for measurements of light intensity, the two labs should perform the measurements under spectra with similar color temperatures.

Interestingly, under the CCT 6000 K spectrum, the PCE is actually reduced for all cases. The GaInP cell shows a PCE of 26.8 % under the 3000 K spectrum but shows a reduced PCE of 25.9 % under the 6000 K spectrum. Likewise, the GaAs1 cell shows a reduction in PCE from 22.8 % under 3000 K CCT to 20.6 % under 6000 K CCT. This finding is in spite of the fact that the electrical performance parameters, including the maximum power point, P_{m}, are actually higher under the 6000 K spectrum than under the 3000 K spectrum. The higher CCT spectra pack more of their optical power in the 400 nm to 500 nm regime but the solar cells have lower spectral responsivities (or external quantum efficiencies) in that regime and are unable to tap into this extra power efficiently; therefore, they show a reduction in PCE.

From among the cells tested here, the GaAs2 device truly outperforms the other cells, including the GaAs1 cell by a significant margin. This cell (by Alta Devices [21], [22]) shows a PCE of ≈ 35 % under the 3000 K CCT reference spectrum. Compared to the GaAs1 cell, the GaAs2 cell enjoys a higher V_{oc} and a larger fill factor (83 % vs 65 %, respectively), directly leading to the higher efficiency observed here. One important factor contributing to the GaAs2 cell's outstanding performance appears to be related to a high external luminescent yield, pushing the V_{oc} up closer to its theoretical limit [23], [24].

IV. Conclusion

We outlined a reference cell-based method for performing low-irradiance *I-V* curve measurements on PV devices under indoor ambient lighting. The reference solar cell was calibrated under a carefully chosen reference irradiance and was used to set an effective irradiance ratio during the measurement. Our results confirm a foregone conclusion in the research community that the spectral composition of the light source does indeed affect the performance of the solar cells. Our methodology and measurement approach clearly explain why this dependence exists and lead the way for developing standards to address these measurement challenges.

Acknowledgement

The author would like to thank John Roller, formerly of NIST, for programming the data acquisition system and Cameron Miller of NIST for graciously making available the CCT computational worksheets to the author. He also thanks Aarohi Vijh of Alta Devices and Yu-Tai Li of ITRI for providing test samples for these measurements.

[1] R. Haight, W. Haensch, and D. Friedman, "Solar-powering the Internet of Things," Science, vol. 353, pp. 124–125, Jul. 2016.

[2] A. Nasiri, S. A. Zabalawi, and G. Mandic, "Indoor Power Harvesting Using Photovoltaic Cells for Low-Power Applications," IEEE Trans. Ind. Electron., vol. 56, pp. 4502–4509, Nov. 2009.

[3] H. Shao, C. Tsui, and W.-H. Ki, "The Design of a Micro Power Management System for Applications Using Photovoltaic Cells With the Maximum Output Power Control," IEEE Trans. Very Large Scale Integr. Syst., vol. 17, pp. 1138–1142, Aug. 2009.

[4] I. Mathews, P. J. King, F. Stafford, and R. Frizzell, "Performance of III–V Solar Cells as Indoor Light Energy Harvesters," IEEE J. Photovoltaics, vol. 6, pp. 230–235, Jan. 2016.

[5] A. S. Teran et al., "Energy Harvesting for GaAs Photovoltaics Under Low-Flux Indoor Lighting Conditions," IEEE Trans. Electron Devices, vol. 63, pp. 2820–2825, Jul. 2016.

[6] P. Vincent et al., "Indoor-type photovoltaics with organic solar cells through optimal design," Dye. Pigment., vol. 159, pp. 306–313, Dec. 2018.

[7] M. Freitag et al., "Dye-sensitized solar cells for efficient power generation under ambient lighting," Nat. Photonics, vol. 11, pp. 372–378, Jun. 2017.

[8] H. K. H. Lee et al., "Organic photovoltaic cells – promising indoor light harvesters for self-sustainable electronics," J. Mater. Chem. A, vol. 6, pp. 5618–5626, 2018.

[9] S. Park et al., "Self-powered ultra-flexible electronics via nano-grating-patterned organic photovoltaics," Nature, vol. 561, pp. 516–521, Sep. 2018.

[10] J. Dagar, S. Castro-Hermosa, G. Lucarelli, F. Cacialli, and T. M. Brown, "Highly efficient perovskite solar cells for light harvesting under indoor illumination via solution processed SnO2/MgO composite electron transport layers," Nano Energy, vol. 49, pp. 290–299, Jul. 2018.

[11] M. Masoudinejad, J. Emmerich, D. Kossmann, A. Riesner, M. Roidl, and M. ten Hompel, "Development of a measurement platform for indoor photovoltaic energy harvesting in materials handling applications," in IREC2015 The Sixth International Renewable Energy Congress, pp. 1–6, 2015,

[12] I. Mathews, G. Kelly, P. J. King, and R. Frizzell, "GaAs solar cells for Indoor Light Harvesting," IEEE 40th Photovoltaic Specialist Conference (PVSC), pp. 0510–0513, 2014.

[13] M. Rasheduzzaman, P. B. Pillai, A. N. C. Mendoza, and M. M. De Souza, "A study of the performance of solar cells for indoor autonomous wireless sensors," 10th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), pp. 1–6, 2016,

[14] H. W. Yoon and C. Gibson, "Spectral Irradiance Calibrations," NIST Spec. Publ., vol. 250–89, 2011.

[15] J. Roller and B. H. Hamadani, "Reconciling LED and monochromator-based measurements of spectral responsivity in solar cells," Appl. Opt., vol. 58, p. 6173, Aug. 2019.

[16] B. H. Hamadani, J. Roller, B. Dougherty, and H. W. Yoon, "Versatile light-emitting-diode-based spectral response measurement system for photovoltaic device characterization," Appl. Opt., vol. 51, p. 4469, Jul. 2012.

[17] B. H. Hamadani, A. Shore, J. Roller, H. W. Yoon, and M. Campanelli, "Non-linearity measurements of solar cells with an LED-based combinatorial flux addition method," Metrologia, vol. 53, pp. 76–85, Feb. 2016.

[18] C. C. Miller, Y. Zong, and Y. Ohno, "LED photometric calibrations at the National Institute of Standards and Technology and future measurement needs of LEDs," in Fourth International Conference on Solid State Lighting, 2004, vol. 5530, pp. 69, October 2004

[19] T. M. Goodman, T. Bergen, P. Blattner, Y. Ohno, J. Schanda, and T. Uchida, "The Use of Terms and Units in Photometry - Implementation of the CIE System for Mesopic Photometry," Int. Comm. Illum., vol. CIE TN004:, 2016.

[20] M. B. Campanelli and B. H. Hamadani, "Calibration of a single-diode performance model without a short-circuit temperature coefficient," Energy Sci. Eng., vol. 6, pp. 222–238, Aug. 2018.

[21] "Certain commercial equipment, instruments, software, 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.." .

[22] E. Yablonovitch, O. D. Miller, and S. R. Kurtz, "The opto-electronic physics that broke the efficiency limit in solar cells," in 2012 38th IEEE Photovoltaic Specialists Conference, 2012, vol. 3, pp. 001556–001559.

[23] O. D. Miller, E. Yablonovitch, and S. R. Kurtz, "Strong Internal and External Luminescence as Solar Cells Approach the Shockley–Queisser Limit," IEEE J. Photovoltaics, vol. 2, pp. 303–311, Jul. 2012.

[24] G. Jarosz, R. Marczyński, and R. Signerski, "Effect of band gap on power conversion efficiency of single-junction semiconductor photovoltaic cells under white light phosphor-based LED illumination," Mater. Sci. Semicond. Process., vol. 107, September 2019, p. 104812, Mar. 2020.

**Proceedings of the ASME 2020 International Mechanical Engineering Congress and Expo
IMECE 2020
November 15-19, 2020, Portland, USA**

IMECE2020- 23180

CAMERA-BASED COAXIAL MELT POOL MONITORING DATA REGISTRATION FOR LASER POWDER BED FUSION ADDITIVE MANUFACTURING

Yan Lu[1*], Zhuo Yang[2], Jaehyuk Kim[3], Hyunbo Cho[3,] and Ho Yeung[1]

1 National Institute of Standards and Technology
Engineering Laboratory
Gaithersburg, MD 20899
Email: [yan.lu, ho.yeung]@nist.gov

2 University of Massachusetts Amherst
Department of Mechanical and Industrial Engineering
Amherst, MA 01003
Email: [zhuoyang]@engin.umass.edu]

3 Pohang University of Science and Technology
Department of Industrial and Management Engineering
Pohang, Gyeongbuk 37673, Republic of Korea
Email: [jaehyuk.kim, hcho]@postech.ac.kr

ABSTRACT

The quality of powder bed fusion (PBF) built parts is highly correlated to the melt pool characteristics. Camera-based coaxial melt pool monitoring (MPM) is widely applied today because it provides high-resolution monitoring on the time and length scales necessary for deep PBF process understanding, in-process defect detection, and real-time control. For such functions, MPM data has to be registered correctly to a well-defined coordinate system. This paper presents methods for camera-based coaxial melt pool monitoring (MPM) data registration using the build volume coordinate system defined in ISO/ASTM52921, for both open architecture AM systems and 3rd party MPM augmented closed commercial systems. Uncertainties are evaluated for the proposed methods and case studies provided to demonstrate the effectiveness of the methods.

Keywords: Additive manufacturing, Data registration, Machine learning, Melt pool monitoring, Powder bed fusion,

1. INTRODUCTION

Laser powder bed fusion (LPBF) is a type of additive manufacturing (AM) that uses a laser beam to melt and fuse material powder layer by layer. Metal LPBF processes involve the spreading of a thin layer of metal powder followed by exposure to high-intensity laser energy directed in scanned trajectories defined by digital models. Both shapes and properties of a material are formed out of the repetitive processes. The metal powder bed fusion process is complicated and stochastic by nature, involving multiple physical phenomena: heat absorption, melt pool formation, solidification, and even re-melting and re-solidification [1]. Investigating melt pool behavior, in both temporal and spatial domains, plays a critical role in understanding and controlling the physical phenomena [2]. However, melt pool behavior is particularly difficult to monitor because of the small size, roughly 50 µm to 250 µm wide, and rapid change, with melting and cooling occurring over approximately 10 µs to 100 µs [3].

Two types of melt pool monitoring (MPM) systems are reported: camera-based and photodetector-based. The former includes both coaxial and off-axial setups of cameras, while the latter measures radiative emission using a photodiode or a pyrometer. This paper focuses on coaxial camera-based melt pool monitoring systems. This type of monitoring can generate high-resolution images, with the melt pool remaining nominally stationary within the field of view, at a high sampling rate. It provides an attractive solution to monitor melt pool for deep

Lu, Yan; Yang, Zhuo; Yeung, Ho; Kim, Jaehyuk; Cho, Hyunbo. "Camera-Based Coaxial Melt Pool Monitoring Data Registration for Laser Powder Bed Fusion Additive Manufacturing." Presented at International Mechanical Engineering Congress and Exposition (IMECE 2020). November 16, 2020 - November 19, 2020.

process understanding, in-process defect detection, and real-time control [4].

Studies have found that melt pool characteristics, including both temperature and geometry, are highly correlated to as-built part structure and properties [5]. Some other studies aim to identify various factors that affect melt pool formation. The factors include processing parameters such as laser power and scan velocity, as well as environmental parameters such as chamber temperature and humidity [6].

Both types of studies aim to derive the relationships between metal PBF process parameters, melt pool characteristics, and material structure and properties. A large volume of design data, in-process control data, melt pool measurement data, as well as post structure inspection data, are collected and analyzed to identify or validate such relationships. These data are generated from various measurement devices or software systems at different AM development stages and hence represented in different spatial reference coordinate systems. For example, AM design data are based on abstract object space models, usually with a reference plane and the build orientation defined. The process controls are prescribed using the machine reference frame, defined by the galvo scanner's positioning of the laser in LPBF, while the actual laser beam positions can be sensed using encoders. Non-Destructive Evaluation (NDE) based post-inspection data has its own reference frames depending on the part location and setup, as well as the scanning configurations. To have a meaningful correlation of melt pool characteristics to process controls or post inspected structure requires transforming all the data into a neutral/standard reference frame or coordinate system. This spatial alignment process is called data registration. Melt pool measurement data registration is a process that assigns every melt pool observation to a geometric location in a selected coordinate system and maps every pixel of each melt pool image to the right location. The latter is necessary to quantify melt pool size and shape as well as the spatter and plume phenomena for part defect prediction.

Based on ISO 20005 [7], data registration is defined as a process of transforming different sets of data into one coordinate system [7]. This is a necessary step for multi-sensor fusion and collaborative information processing. Sensor modeling and calibration are the basis for sensing data registration. For camera-based sensing systems, various camera models can be built to map an image pixel to a real-world coordinate. The mathematical relationship can be calculated using the camera's intrinsic and extrinsic parameters [8]. If a camera's optical setup is complicated, e.g., a coaxial melt pool monitoring imaging system, the camera model can be obtained through a homography mapping using calibration data [9]. In medical applications, image data registration is defined beyond pixel mapping, usually referring to feature-based methods to find correspondence between image features such as lines and contours [10]. A deep-learning-based data registration method was also proposed to allow more complex data set to be aligned, like conceptual models as well as 3D datasets [11].

In the AM domain, multiple studies are reported on camera-based coaxial MPM data collection and calibration. Zhirnov et al. reported a study to locate the position of the laser spot with respect to the imaged melt pool [12]. Research from NIST has resulted in several complete data sets with both control commands and process monitoring data, including images captured by a coaxial high-speed camera [13, 26 and 27]. The studies describe a melt pool monitoring system for an open architecture AM system, and data sets generated. However, neither of them addressed a systematic way to register the melt pool images to allow for process-structure-property correlation. A closed AM system with melt pool monitoring built on top of a commercial AM machine was described in [14]. Their data analysis and derived predictive model were limited to layerwise statistics because the melt pool images were acquired by an external MPM system, and the images were not registered to the local positions. No research work has been reported on data registration for this type of melt pool monitoring system, which are 3rd party addons to commercial AM systems.

This paper intends to fill the gap by introducing data registration methods for camera-based coaxial melt pool monitoring of powder bed fusion AM processes. First, we present a generic melt pool monitoring image data registration procedure with a formal problem definition. Second, we present data registration algorithms for both open architecture AM systems and closed AM systems with 3rd part MPM additions. For both cases, each image frame is first registered against its corresponding laser beam spot position. Then individual pixel registration is performed based on the relative distance of a pixel to the melt pool center. For a closed AM system, the data acquisition of the MPM system is not synchronized with the AM machine positioning system. Hence the laser beam position for each melt pool image frame is not available and has to be estimated. In this study, we explored the most advanced machine learning algorithms and solved the closed-system melt pool position estimation problem. Uncertainties are evaluated for the proposed methods, and case studies are provided to demonstrate the effectiveness of the methods.

The paper is organized as the following. Section 2 presents the build volume coordinate systems defined by ISO/ASTM52921 [28] and formulates the data registration problem. Section 3 presents the data registration method for open architecture systems. Section 4 describes the camera-based coaxial MPM data registration for commercial AM systems with augmented 3rd party process monitoring. Section 5 summarizes the paper with discussions and future work.

2. MPM DATA REGISTRATION OVERVIEW

In this section, we introduce the reference coordinate system used for in-situ data registration, formulate the camera-based coaxial MPM data registration into two problems, and define a general procedure to solve the problems. We also define two typical scenarios for solving the problem: data registration for open architecture AM systems and data registration for closed AM systems.

2.1 A Reference Coordinate System

There are multiple coordinate system types defined in manufacturing practices, for example, machine coordinate system, workpiece coordinate system, and tool coordinate system. All of them are "right-handed" three-dimensional Cartesian coordinate systems well defined by ISO 841 [15] and ISO 2806 [16]. For AM, ISO/ASTM 52921:2013(E) defines a build platform coordinate system. The build facing surface of the build platform is defined as the X-Y plane, and the center of this surface is the origin, named *build volume origin* in Figure 1. The build volume origin is different from the machine origin, as defined by the original equipment manufacturer, usually named as *Machine Zero*. In addition, this definition may not coincide with the galvo-based coordinate origin. The X-axis is defined parallel to the front of the machine and is horizontal and parallel with one of the edges of the build platform. The Y-axis runs perpendicular to the X-axis and parallels to the other edges of the build platform. The Z-axis runs normal to the building facing surface, pointing to the build orientation. For our study, instead of using continuous values for the Z-axis, we define Z-coordinates as a set of discrete layer numbers *{1,2,…N},* from the first layer to the last layer N.

Since the build platform coordinate system has a priority for AM applications, we adopt this definition as our reference coordinate system for melt pool monitoring data registration.

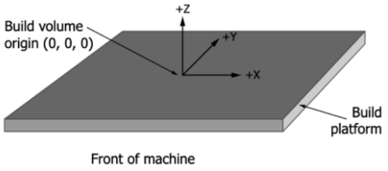


Figure 1: A reference coordinate system defined based on ISO/ASTM 52921

With the defined coordinate system, any point on the build surface for a specific layer process can be represented as *(x, y, n),* where *x* and *y* are confined by the build volume, and *n* is limited by the maximum layer number for any given layer thickness settings.

2.2 MPM Data Registration Problems and Procedure

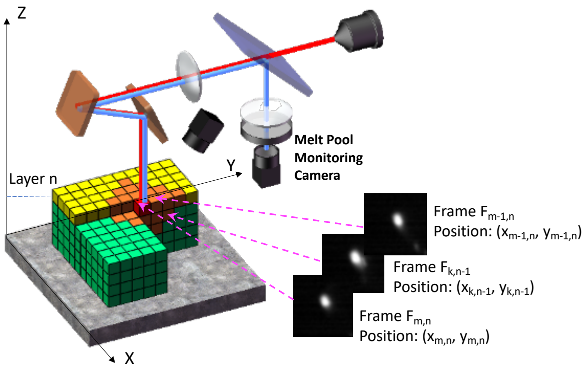
The data generated from a camera-based coaxial MPM system includes layerwise time series of melt pool images, which can be represented as $F_{m,\,n}$, where *n* represents the corresponding layer number when the series of images are taken, and *m* represents the index of the frame in that time series. Assuming the pixel window size of the camera is *W*x*H*, then any given pixel *(i, j)* of the image frame $F_{m,\,n}$ can be represented as $(Px_i, Py_j)_{m,\,n}$, where $i=1\ldots W$ and $j=1\ldots H$.

Thus, a camera-based coaxial MPM data registration method is needed to solve two problems:

Problem 1: For any melt pool image $F_{m,\,n}$, find the corresponding laser beam location $(x_{m,\,n}, y_{m,\,n})$

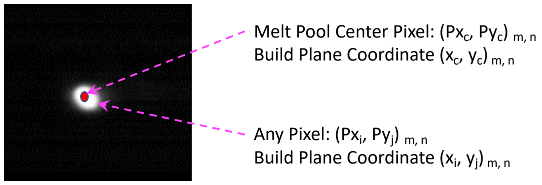
Problem 2: For any pixel *(i, j)* of the $F_{m,\,n}$, named $(Px_i, Py_j)_{m,n}$, find the corresponding build volume coordinate location $(x_i, y_j)_{m,\,n}$.

Figure 2(a) and (b) illustrate the two problems, respectively. A solution to Problem 1 is required for data visualization and qualitative data fusion, for example, classifying and clustering melt pools based on the size or shape and correlating the results with porosity measurements. Problem 2 has to be solved if one needs to measure the absolute size of a melt pool or locate and quantify the phenomena of spatter and plume. More process features can be extracted based on Problem 2 solutions.



Z
Y
X
Layer n
Melt Pool Monitoring Camera
Frame $F_{m-1,n}$
Position: $(x_{m-1,n}, y_{m-1,n})$
Frame $F_{k,n-1}$
Position: $(x_{k,n-1}, y_{k,n-1})$
Frame $F_{m,n}$
Position: $(x_{m,n}, y_{m,n})$

(a)



Melt Pool Center Pixel: $(Px_c, Py_c)_{m,\,n}$
Build Plane Coordinate $(x_c, y_c)_{m,\,n}$
Any Pixel: $(Px_i, Py_j)_{m,\,n}$
Build Plane Coordinate $(x_i, y_j)_{m,\,n}$

(b)

Figure 2: Camera-Based Coaxial MPM Data Registration Problems - Registering an image frame (a) Individual image registration (b)

With the problems formulated aforementioned, a general procedure for the camera-based coaxial MPM data registration can be identified, including 5 Steps.

Step 1: Registering Sensor Information – during this phase, all the sensor-related information required by the data registration algorithms should be collected. In general, the sensor information can be classified into three categories: sensor parameters, reference frame information, and/or camera calibration data. How the optical path is designed decides how to calculate the spatial resolution for each image.

Step 2: Acquiring, Cleaning, Archiving, and Indexing images – this phase focuses on collecting images, treating missing data or contaminated data, storing and labeling melt pool images frame by frame, layer by layer, using the index mechanism defined in Problem 1.

Step 3: Locating Image Frame – during this phase, a solution to Problem 1 is found. Every image frame for each layer is registered to a location on the build surface based on the coordinate system defined by ASTM.

Step 4: Locating Melt Pool Center – melt pool center is defined as the pixel within the MPM image that coincides with the center of the laser spot. The laser spot center may not coincide with the centroid of the melt pool image [12].

Step 5: Registering Melt Pool Image – this step assigns a build surface position to any pixel of a melt pool image, based on the relative pixel distance to the melt pool center and the pixel spatial resolution.

Steps 1 and 2 are relatively trivial. However, data preprocessing is a very important step. Images acquired should be scanned for missing frames or corrupted melt pool measurements. Melt pool no-shows on part contours or hatch regions indicate faults. Missing frames by camera or data transfer errors should be marked before data registration. Step 3 to 5 are challenging due to the dynamic nature of PBF processes and the complexity of MPM measurement setups, as well as the various latencies characterizing data acquisition systems. This paper focuses on solving the two MPM data registration problems needed for Step 3 to 5.

We address the problems in two different scenarios of AM system settings. Their given conditions are different; hence the solutions are different. In the first scenario, an open AM system is considered. An open system allows full control of laser scans as well as synchronized data acquisitions. For example, the NIST AMMT system uses a real-time controller to set galvo position and laser power as well as trigger acquisition of MPM images [17]. Another example is the GE Open Architecture system, which employs two embedded systems for open control and process monitoring separately [18]. The second scenario is more common where an AM system is built based on a closed AM machine with one or more 3rd party melt pool monitoring systems. The 3rd party melt pool monitoring data acquisition is independent of the AM system motion and laser control [14]. In this case, the associated laser beam center position for every image frame has to be estimated from the melt pool image characteristics.

3. OPEN ARCHITECTURE PBF SYSTEM CO-AXIAL MPM DATA REGISTRATION

3.1 Method Overview

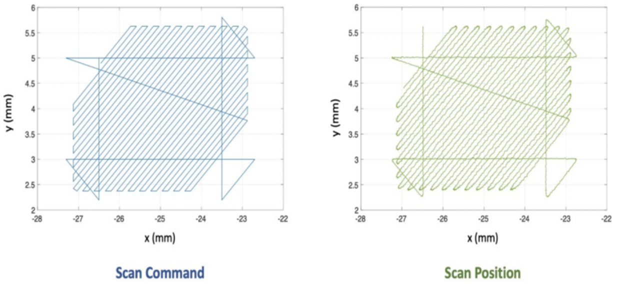
Two types of open architecture AM systems were mentioned in Section 2. First is a custom-built system with full control of AM processes and data collection for melt-pool monitoring [17]. The melt pool data and laser spot positions from this type of system can be aligned using the camera triggers embedded in xy2-100 formatted build commands. The delays between the command signals and measured signals should be calibrated and removed for the temporal alignment. The second type of open architecture system doesn't have synchronized positioning and camera triggering commands to align the smart camera-generated timestamps with the encoder positions acquisition [18]. However, the delay between the two systems is constant, and it can be estimated based on a synchronized laser power measurement. This section illustrates a frame registration method for the first case, while it can be easily simplified for the second case.

Equation (1) shows a temporally aligned data matrix of scan commands, real measurement, and the camera control trigger based on xy2-100 format [19]. The first and last columns are the time stamp and camera control trigger, respectively. The synchronization of these two columns can guarantee the data of each row to be fully aligned. Columns 2 to 4 are the original scan command. Columns 5 to 7 are the measured scan position. Note, for multi-layers data; each layer has its corresponding matrix.

$$\begin{bmatrix} T_1 & x_1^{COM} & y_1^{COM} & P_1^{COM} & x_1^{MEA} & y_1^{MEA} & P_1^{MEA} & Tr_1 \\ T_2 & x_2^{COM} & y_2^{COM} & P_2^{COM} & x_1^{MEA} & y_2^{MEA} & P_2^{MEA} & Tr_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ T_i & x_i^{COM} & y_i^{COM} & P_i^{COM} & x_i^{MEA} & y_i^{MEA} & P_i^{MEA} & Tr_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad (1)$$

Figure 3 shows an example of a time series of scan command and the corresponding real scan position. It is noted that the real scan position is not identical to the scan command

1) The measured position is noisy compared to the command
2) The inertia of a galvo scanner is substantial compared with the sample rate. However, the delay can be measured. [17]



Scan Command

Scan Position

Figure 3: Scan commands vs. real scan positions

3.2 Image Frame Registration Against Laser Spot Position

The synchronized time stamps and camera triggers can be used to register the image frame against the laser spot position in the build platform coordinate system. Each image frame firstly is paired with its trigger command, and the laser position command correspondingly. However, it is preferred to register the melt pool with the real position measurement as it provides the most accurate spatial correlation with the as-built part and the Computer-Aided-Design (CAD) model. In order to perform that, a delay measurement must be conducted during the galvo calibration process. Assuming that the control and position measurement sampling rate is T_s, the delay between the scan command and the scan position is t_{d1}, and the delay between the triggering command and the camera shutter opening is t_{d2}, Column 4-6 in Matrix (1) should be moved up for $(t_{d1}-t_{d2})/T_s$ rows to align actual laser beam positions with melt pool images.

After that step, at T_{rm} of layer n, image frame $F_{m,n}$ would be registered to point $[x_m^{MEA}, y_m^{MEA}]_n$ defined in the ISO/ASTM 52921 build coordinate system.

3.3 MPM Image Pixel Registration

In Section 3.2, each image frame is registered as a geometric point with no area, which completes Step 3, as defined in Section 2.2. This registered information can help the user to investigate the general melt pool conditions over the part. However, the image frame registration cannot provide enough details such as melt pool overlapping between tracks or a prediction of lack of fusion defects. To achieve this goal, each MPM coaxial image needs to be registered to the build platform coordinate system to provide a pixel-to-build plane projection.

For coaxial image systems, the imager is optically aligned with the laser beam such that the field of view moves with the laser spot. Consequently, nominal stationary melt pool images are obtained. The image window size is unchanged during the scan process. However, the pixel size could vary when the laser moves from the center of the build plane to the edge of the build platform, due to the distortion and stretching caused by perspective distortion from the location-dependent, non-orthogonal viewing angles to the build plane [20, 21]. Research shows the real laser spot size is elongated when the laser beam shot on the build plate, not perpendicularly [21]. The pixel size calibration over the build platform is requested before the registration process.

To simplify this issue, the build platform can be divided into $k \times k$ grids, where each grid should cover a maximum area with minimum deflection angle change. Larger build platform results in larger deflection angle change, thus demanding a larger *k*. Since each grid represents an area with a very small deflection angle range, it is reasonable to assume the pixel size is the same within one grid. If the part placed on the intersection area of multiple grids, the calibration would use the average size of the grids. It is assumed that $N_{xp}\ \mu m \times N_{yq}\ \mu m$ is the calibrated real pixel size for grid *(p, q)*, where x and y represent the pixel stretching direction.

Figure 4 shows an example of a gridded build plane, where the red area represents the hypothetical pixel distortion in different grids for illustration purposes. The specified spatial resolution is located in the center grid. The stretching ratio depends on the distance from the current grid to the center grid and the corresponding laser/camera view angle . The gray rectangle represents a $3\,mm \times 2\,mm$ test part placed at this location, referring to the example in Section 3.4.

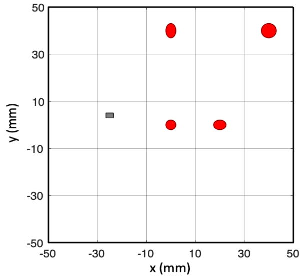


Figure 4. Pixel calibration grid for a build plane

Normally, both melt pools and laser spots cover more than one pixel area. The registration of a melt pool image assumes the center of a melt pool is perfectly mapped to the center of the laser spot. In this case, the center of the melt pool in an image frame should be identified and would be registered to the laser spot position first. The pixel located at the center of the area with the highest temperature is assumed to be the laser spot center [12]. After the center pixel$(P_{x_c}, P_{y_c})_{m,n}$ for $F_{m,n}$ is located, that pixel is registered to the position $[x_m^{MEA}, y_m^{MEA}]$, which was identified in Step 3. The center pixel then can be used as the reference to register the remaining pixels. To register any pixel (i, j) of that image, the grid should be identified for this image frame, as (p, q). Using a lookup table, the pixel size is found as $N_{xp}\ \mu m \times N_{yq}\ \mu m$. Then, $(P_{xi}, P_{yj})_{m,n}$, the pixel at (i, j) in the image, should be registered at

$$[x_m^{MEA} + (P_{xi} - P_{xc})N_{xp},\ y_m^{MEA} + (P_{yi} - P_{yc})N_{yq}] \qquad (2)$$

3.4 An Image Registration Example of AMMT

This section presents an example of the image frame and pixel registration based on the AMMT open platform [17]. The experiment builds a $3\,mm \times 2\,mm$ single layer part from IN625 powder located as marked in Figure 4. The experiment uses an orthogonal skywrite scan pattern with a constant laser power 195 W and a scan speed 800 mm/s. The part was built by 24 646 timestamps with 10 µs time interval. The coaxial camera is triggered every 50 µs to collect 1497 image frames. There is no laser input and image collected at the overshooting area. During the experiment, the encoder measures the real scan position.

Table 1 lists the fully aligned data of the first 6 time stamps. It roughly shows a 40 µs signal delay and a 0.05% position error. As shown in the table, the first image frame was captured at the beginning of the scan at the start point [-26.487 mm, 4.9985 mm].

Table 1. XYPT command for galvo, laser and camera control

Time Stamp	Scan Commands			Real Measurement			Camera Control
T_i (µs)	X_{COM} (mm)	Y_{COM} (mm)	P_{COM} (W)	X_{MEA} (mm)	Y_{MEA} (mm)	P_{MEA} (W)	Tr
0	-26.50	5.000	195	-26.487	4.9985	195	2
10	-26.50	4.992	195	-26.484	4.9985	195	0
20	-26.50	4.984	195	-26.484	4.9985	195	0

30	-26.50	4.976	195	-26.484	4.9985	195	0
40	-26.50	4.968	195	-26.484	4.9870	195	0
50	-26.50	4.960	195	-26.489	4.9704	195	2

Figure 5 shows the image frame registration result. The top plot shows the scan path and trigger position of the scan commands. The bottom plot shows the real position measured by the encoder. Both plots use the build platform coordinate system. Three sample coaxial images are marked in red, blue, and green frame. The corresponding timestamp is 28260 µs, 82620 µs, and 109410 µs, respectively. Both plots mark the trigger position by solid black dots. The figure extracts three example frames and marks their registered point on the plots. Each colored frame is registered to the point of the same color.

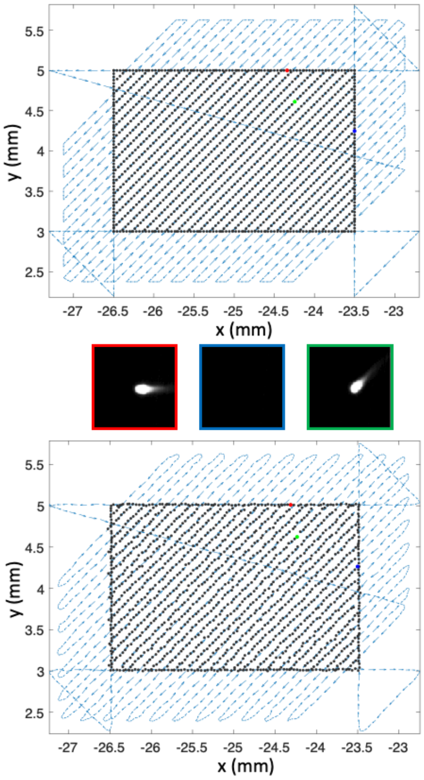


Figure 5. Image frame registration against scan command (top) and real measurement (bottom)

For AMMT, the distance from the build platform (100 mm x 100 mm) to the galvo is 500 mm. The laser beam deflection range is thus between -5.7 Deg to 5.7 Deg. According to the pixel size of AMMT coaxial image, 8 µm x 8 µm, the maximum distorted pixel is 8.04 µm x 8.04 µm. As a result, this example divides the build plane into a 5x5 grid, where the pixel size within one grid is assumed to be the same. For this build, the pixel would keep the original size in the y-direction and stretch in the x-direction. The calibrated pixel size is 8.0064 µm x 8 µm.

Figure 6 visualizes a registered image on the build plane. The original coaxial image is shown on the right. From the image frame registration work, it is known that this frame is registered at [-24.361 mm, 3.664 mm]. The calibrated temperature image is shown under the original. The dark red area has a temperature of over 2200°C. The geometrical center of this area is assumed to be the laser spot center. The center pixel is marked in black color and located at row 60 column 70. The center pixel is registered to the [-24.361 mm, 3.664 mm]. The pixel represents an area of $\begin{bmatrix} -24.365\text{ mm} & -24.357\text{ mm} \\ 3.660\text{ mm} & 3.668\text{ mm} \end{bmatrix}$ in the plot of build platform coordinate system. The left figure shows the build platform coordinate system after registered all the pixels of this frame.

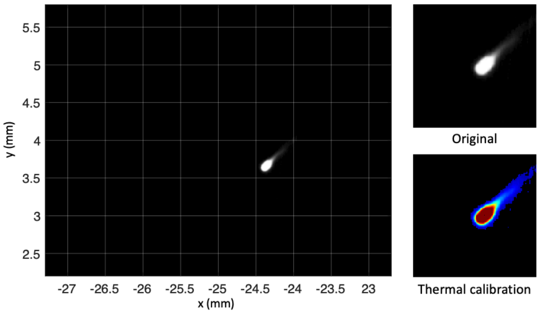


Figure 6: Coaxial image pixel registration example

Both an image frame and pixels are registered under uncertainties. The major uncertainty resources of the image frame registration are the signal delay and position variation. Figure 7 (a) and (b) plot the error distribution of the scan in x and y directions, respectively. The error is evenly distributed on both axes over the entire part. When registering an image frame using the related scan command, this uncertainty would be persisted. In this study, the real measured position is assumed to be the ground truth. However, it may include uncertainties from the measurement.

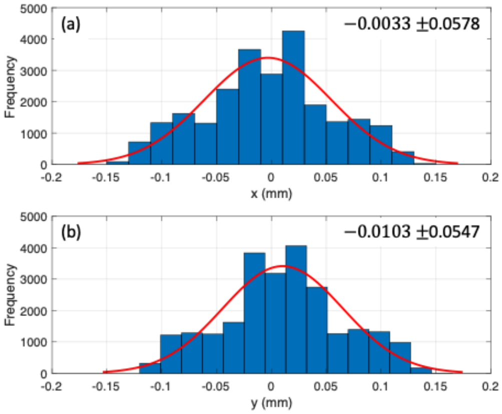


Figure 7. Statistical analysis for melt pool image registration. (a) and (b) are the distribution of error between real measurement and scan commands.

Image pixel registration combines the uncertainties from the previous step and the uncertainty initiated from the coaxial images. First, the pixels registration refers to the frame position. The error of the position can be accumulated from pixel to pixel. Second, the pixel position depends on the laser spot found on the coaxial image, which is assumed to be closed to the image center. In this example, the image ($960\ \mu m \times 960\ \mu m$) center is at [480 µm, 480 µm]. However, the laser spot center is randomly distributed in a $100\ \mu m \times 100\ \mu m$ area at $[534 \pm 25, 454 \pm 23]$. The distribution is shown in Figure 8 as a heat map. The uncertainties analysis may be used for the optical and camera system adjustment.

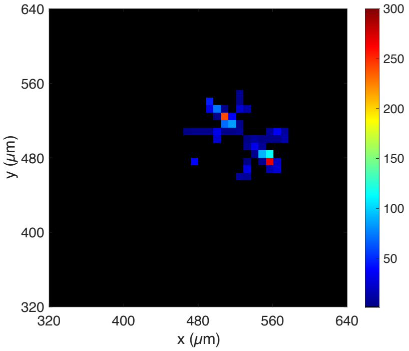


Figure 8. Heatmap of the center pixel frequency

4. CLOSED PBF SYSTEM CO-AXIAL MPM DATA REGISTRATION

4.1 Method overview

A closed AM system includes a commercial PBF machine as a base and many 3rd party addons such as in-situ monitoring systems. Because of the proprietary nature, most commercial AM machines don't give away the exact scan commands. Hence addon MPM system data registration is challenging because a 3rd party data acquisition is independent of the AM machine process control. And laser spot real position measurement during a build is seldom available to the users without breaking the warranty of the commercial machine. Hence the existing studies of melt pool characteristics and their correlations to other AM data are limited to layerwise global statistics because the laser spot positions for individual melt pool images are not available [14]. For this type of MPM systems, the real scan position for each melt pool has to be estimated based on machine-provided process settings and melt pool images themselves. This corresponds to Problem 1 for MPM data registration. Since Problem 2 for closed systems can be solved in a similar way in Section 3, it will not be discussed in this section.

As shown in Figure 9, a typical scan layer of a part is divided into contour region, hatching region, and skywriting region based on exposure strategy (laser power (P) and scan speed (v)) [22]. The contour region is an exposure area that creates the contour of a part, and a relatively small scale melt pool is formed due to low laser power and high scan speed. The hatching region is an exposure area that creates a core inside the contour of a part, usually using a higher laser power and a lower scan speed. Relatively large scale melt pools are formed in these regions. Figure 9 shows a stripe pattern for infill, one of the most frequently used scan strategies by commercial AM systems. The skywriting region is an unexposed area for scanner acceleration or de-acceleration in order to generate uniform energy density (P/v) in the hatch region. In this region, the melt pool is not formed because the laser is switched off.

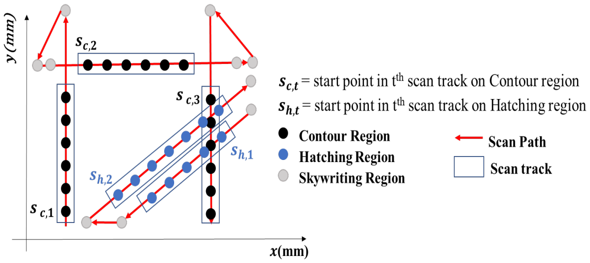


Figure 9: An example scan profile and scan regions

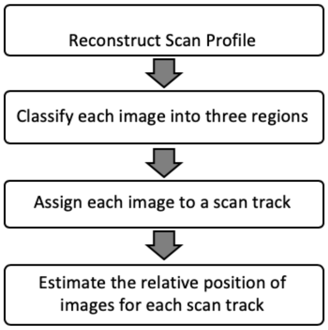

Reconstruct Scan Profile
Classify each image into three regions
Assign each image to a scan track
Estimate the relative position of images for each scan track

Figure 10: MPM data registration for closed systems

Figure 10 shows the overall workflow of the proposed MPM data registration method for closed AM systems based on the scan region definition.

1) A scan profile should be reconstructed from a process setting. Without losing generality, part contours, laser power for the contour (P_c), hatch pattern and direction (α), hatch distance (h_d), hatch laser power (P_h), and scanning speed (v) are given. With the process setting, a scan profile can be obtained, which is similar to Figure 9, including the contour(s) and hatch tracks.

2) Layerwise organized time series of melt pool images are classified into various regions. As shown in Figure 9, the melt pool images in the three regions are plotted as grey, black, and blue, respectively, in the figure. The details of the image classification will be discussed in the next sub-section.

3) The transient images are identified, e.g., from the skywriting region with "no melt pool" to the hatch region "with melt pool" or the other way around. These images correspond to the Start and End points of each scan tracks in a region, as grouped in the light blue rectangles in Figure 9. The start of a contour track is marked as $S_{c,t,}$ and that of the end point is marked as $E_{c,t}$. Similarly, the start and end points of the hatch tracks are marked

as $S_{h,t,}$ and $E_{h,t}$. With the Start and End points identified, the images in between can be assigned to individual tracks.

4) Laser spot positions are assigned to the images, as shown in Figure 11. The corner points on the contour are assigned to the Start and End points on the contour tracks. The cross points of the contour and the hatch tracks identified in Step 1 are assigned to the Start and End points for the hatch tracks. For the images between the Start and End points on a trach, their positions can be obtained through linear extrapolation, from the position of previous melt pool image frame, the scan speed and sampling time shown in the equations below

$$\boldsymbol{x_{m+1,n}} = \boldsymbol{x_{m,n} + t \times v \times cos(\alpha)} \tag{3}$$
$$\boldsymbol{y_{m+1,n}} = \boldsymbol{y_{m,n} + t \times v \times sin(\alpha)} \tag{4}$$

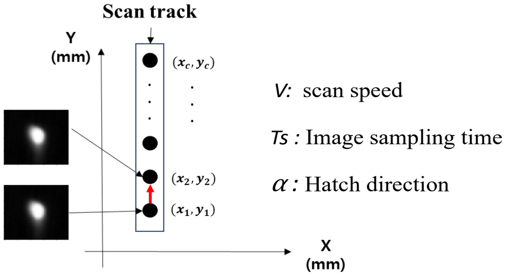


Figure 11: On-track image registration

4.2 Melt Pool Image Classification

The melt pool image can be classified based on the region because each region has a different exposure setting that determines the characteristics of the melt pool. In the skywriting region, no melt pool is taken in any images because the laser is turned off there. All pixel values of the image are zero and so can be easily distinguished. However, melt pool images in the regions with laser exposures, e.g., the contour region and the hatching region, are difficult to be separated by the pixel values due to various shapes of the melt pool, such as a long tail.

Because of differences with manufacturing conditions between 3rd parties, the rule-based methods of feature extraction, such as edge detection [25] and classification, cannot be used. To solve this problem, advanced machine learning algorithms are explored. To extract features of a melt pool image and classify whether the image is on a contour region or a hatching region, PCA and logistic regression are used respectively. PCA is a technique that finds latent variables orthogonal to each other while preserving the variance of data as much as possible and transforms variables from high-dimensional spaces into low-dimensional spaces without linear correlation [23]. Logistic regression is the most frequently used statistical model that obtains the probability value $P(\hat{Y})$ which is mapped to two classes (0 or 1) using a linear combination of input variables$x_1, x_2, \cdots, x_n$.(Equation (5)) [24]. Therefore, PCA is used to reduce high-dimensional image data (e.g., 61 pixel x 61 pixel = 3721) to low-dimensional data without linear correlation, and logistic regression is used to classify images into two types using the low-dimensional data as input variables.

$$\hat{Y} = \beta_0 + \beta_0 x_1 + \cdots + \beta_n x_n = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{5}$$

where $\hat{Y}$ is the predicted value, β_0 is the intercept and$\beta_1, \beta_2, \cdots, \beta_n$ are the coefficients which are estimated by maximum-likelihood. To deal with binary-classification, the real value of Y is transformed into the probability of Y occurrence with Equation 6.

$$P(\hat{Y}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \tag{6}$$

4.3 Case study

This section presents a case study of the proposed method for closed-system MPM data registration. This study uses the melt pool images collected from NIST AMMT [17]; however, assuming the scan commands and laser spot positions unavailable. Instead, the scan command is used for algorithm validation. The measured laser spot position is not available for this set of data2

A layer with 2661 melt pool images is selected to validate the closed-system MPM data registration approach. Each image has a size of 128 pixels x 120 pixels at grayscale 0 to ≈255. The coaxial camera is triggered every 500 µs to collect melt pool image frames. To classify whether a melt pool image is on the contour or in the hatching region, a total of 2294 images are identified, with 60 images on the contour and 2134 images in the hatching region. 1720 images are used for training, and 574 are used as the test set for classification accuracy.

Table 1 is the scan profile of the NIST AMMT study [17].

TABLE 1: Scan profile of process parameters from NIST AMMT

Region type	Scan speed (mm/s)	Laser power (W)	Scan & Hatch direction(°)
Contour	900	100	{0,90,180,270}
Hatching	500,800	195	{67,247}
Skywriting	{0,...,900}	0	-

Figure 12 shows example images by region.

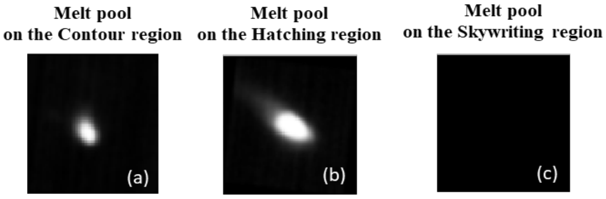

Melt pool on the Contour region
Melt pool on the Hatching region
Melt pool on the Skywriting region
(a)
(b)
(c)

Figure 12: Example of melt pool image frame by region.

With 14 principal components resulted from PCA, Figure 13 shows two classified clusters with PCA and the result from logistic regression. Using the number of principal components of 14, the cumulative explained variance is 98%. With the 14 principal components of the training dataset, the classification accuracy of test sets is 99.65%.

As shown in Figure 13, the classification errors are those hatch track images mistakenly classified as contour images. Rules are built in the algorithm to automatically correct this type of error. The rule finally corrects 4 errors for the case study, which helps to avoid significant position estimation errors.

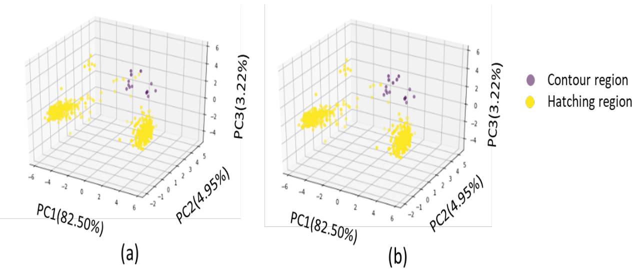


Figure 13: Plots of the logistic regression result in three principal axes. (a) is the result of the test dataset, and (b) is the predicted test dataset.

Figure 14 shows the MPM data registration using laser beam location estimation compared with the one registered against the build command. The top plot shows the registration against scan command, and the bottom plot shows the MPM data registration against the estimated positions using the proposed method. Both plots use the build platform coordinate system. The position error for all frames is less than 1 %.

The registration errors from the proposed laser spot position estimation methods come from 4 areas:

1) The scan profile reconstruction error
2) The laser spot position assigning error. Assigning the contour-hatch track cross point positions to the Start and End images on the hatch track could involve a position error as large as the laser travel distance within one image sampling time, about 400µm, for 500µs sampling time and 800mm/sec scan speed.
3) The algorithm assumes that there is no contour during the regular scan. The wrong assumption can cause significant classification error and translates to the position error.
4) The extrapolation error from nonlinear scan velocity

5. DISCUSSION AND FUTURE WORK

The objective of this work is to develop a generic data registration method and algorithms for camera-based coaxial melt pool monitoring for both open architecture AM systems and closed AM systems with 3rd part MPM additions. For an open AM system, each image taken by the coaxial MPM system is registered against the corresponding laser beam spot position based on synchronized camera trigger signals with the laser spot commands. The alignment between the melt pool images and the real laser positions can be made by getting rid of the delays introduced by the galvo position system and the camera control system. The delays can be measured during the calibration process. The individual image registration can be made based on the melt pool center identification and the calibrated camera pixel resolutions on the build plane. With the melt pool data accurately registered, 3D models can be reconstructed to correlate to the CAD model or post-inspection data. The relationships will help in-situ part quality control and quality prediction. The machine learning-based laser spot position estimation method enables the MPM data registration for closed AM systems, which will benefit both the research labs and production facilities to enhance their integration capability at a lower cost. In the future, we will further improve the data registration method for closed AM systems for more complex scan patterns, by incorporating melt pool direction modeling to the on-track image position estimation. Deep learning methods will be explored for both image clustering and scan direction estimation. Deep dive uncertainty analysis will be conducted as well. For data interoperability and usability, these methods should be turned into standards. This requires us to work with the AM machine vendors and data users to further refine the data

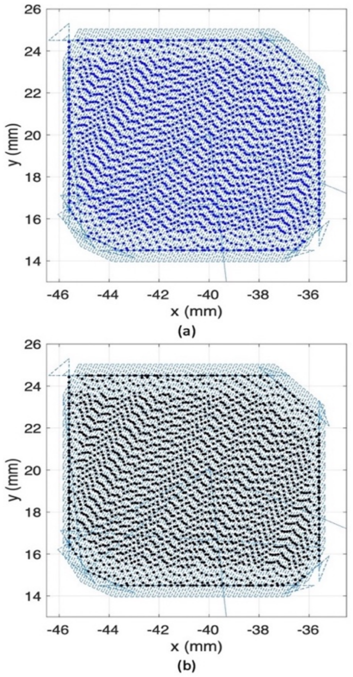


Figure 14: (a) MPM data registration against the scan command, and (b) MPM data registration against the estimated laser spot position

registration methods. This is also one of the directions we are heading.

REFERENCES

[1]. Frazier, W. E., “Metal Additive Manufacturing: A Review”, Journal of Materials Engineering and Performance, (2014) 23:1917–1928, ASM International, DOI: 10.1007/s11665-014-0958-z

[2]. King, W. E., Anderson, A. T., Ferencz, R., 2015, "Laser Powder Bed Fusion Additive Manufacturing of Metals; Physics, Computational, and Materials Challenges," Applied Physics Reviews, 2(4) pp. 041304.

[3]. Fisher, B.A., Lane, B., Yeung, H., and Beuth, J., “Toward determining melt pool quality metrics via coaxial monitoring in laser powder bed fusion”, Manufacturing Letter, Manufacturing Letters 15 (2018) 119–121

[4]. Fox, J., Lane, B., and Yeung, “Measurement of process dynamics through coaxially aligned high speed near-infrared imaging in laser powder bed fusion additive manufacturing, Proc. SPIE 10214, Thermosense: Thermal Infrared Applications XXXIX, 1021407 (5 May 2017); https://doi.org/10.1117/12.2263863.

[5]. Grasso, M., Colosimo, B. M. “Process defects and in situ monitoring methods in metal powder bed fusion: a review,” Meas Sci Technol (2017), 28:044005. doi:10.1088/1361-6501/aa5c4f.

[6]. Yang, Z., Lu, Y., Yeung, H., and Krishnamurty, S., "Investigation of Deep Learning for Real-Time Melt Pool Classification in Additive Manufacturing," *2019* IEEE 15th International Conference on Automation Science and Engineering *(CASE)*, Vancouver, BC, Canada, 2019, pp. 640-647, doi: 10.1109/COASE.2019.8843291.

[7]. ISO 20005, “Information technology — Sensor networks — Services and interfaces supporting collaborative information processing in intelligent sensor networks”, https://www.iso.org

[8]. Hartley, R. and Zisserman, A., “Multiple View Geometry in Computer Vision”. Cambridge University Press, 2003

[9]. Chum, O., Pajdla, T., and Sturmb, P., “The geometric error for homographies”, Computer Vision and Image Understanding 97 (2005) 86–102

[10]. Brown, L. G., “A survey of image registration techniques”, ACM Computing Surveys archive, volume 24, issue 4, December 1992), pages 325 - 376

[11]. Villena-Martineza, V., Opreaa, S., Saval-Calvoa, M., Azorin-Lopeza, J., Fuster-Guilloa, A., and Fisherb, R. B., “When Deep Learning Meets Data Alignment: A Review on Deep Registration Networks (DRNs)”, https://arxiv.org/pdf/2003.03167.pdf, Accessed on May 7, 2020.

[12]. Zhirnov, I., Mekhontsev, S., Lane, B.M., and Grantham, S.E., “Accurate Determination of Laser Spot Position during Laser Powder Bed Fusion Process Thermography”, manufacturing Letters, Vol. 23 (2020) 49-52

[13]. Lane, B., and Yeung, Ho, “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, Volume 124, Article No. 124033 (2019) https://doi.org/10.6028/jres.124.033

[14]. Yang, H. C., Adnan, M., Huang, C. H., Cheng, F. T., Lo, Y. L., and Hsu, C., H., “An Intelligent Metrology Architecture with AVM for Metal Additive Manufacturing”, IEEE Robotics And Automation Letters, Vol. 4, Issue 3, 2020.

[15]. ISO 841:2001, “ Industrial automation systems and integration — Numerical control of machines — Coordinate system and motion nomenclature”, https://www.iso.org

[16]. ISO 2806:1994, “Industrial automation systems — Numerical control of machines — Vocabulary”, https://www.iso.org

[17]. Lane, B. M., Mekhontsev, S., Grantham, S. E., Vlasea, M., Whiting, J. G., Yeung, H., Fox, J. C., Zarobila, C. J., Neira, J. E., McGlauflin, M. L., Hanssen, L. M., Moylan, S. P., Donmez, A. M., and Rice, J. P., “Design, Developments, and Results from the NIST Additive Manufacturing Metrology Testbed (AMMT)”, Proceedings of the Solid Freeform Fabrication, August 8-10, 2015, Austin, TX

[18]. Carter, W., Tucker, M., Mahony, M., Toledano, D., Butler, R., Roychowdhury, S., Nassar, A. R. , Corbin, D. J., Benedict, M. D., and Hicks, Adam S., “An Open-Architecture Multi-Laser Research Platform for Acceleration Of Large-Scale Additive Manufacturing (ALSAM)”, Proceedings of the 30th Annual International, Solid Freeform Fabrication Symposium, August 10-12, 2019, Austin, TX

[19]. Yeung H, Neira J, Lane B, Fox J, Lopez F (2016) Laser path planning and power control strategies for powder bed fusion systems. Proceedings of the 27th Annual International Solid Freeform Fabrication Symposium, Austin, TX. 113–127.

[20]. Li, Yajun. "Beam deflection and scanning by two-mirror and two-axis systems of different architectures: a unified approach." Applied optics 47, no. 32 (2008): 5976-5985.

[21]. Yang, Pei-Ming, Yu-Lung Lo, and Yuan-Hao Chang. "Laser galvanometric scanning system for improved average power uniformity and larger scanning area." Applied optics 55, no. 19 (2016): 5001-5007.

[22]. EoS, PSW software manual, 2008

[23]. Jolliffe, I.T., Principal Component Analysis. Springer-Verlag, 1986.

[24]. Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X., Applied Logistic Regression, p. 1. John Wiley & Sons, third ed., 2013.

[25]. Sampson, R., Lancaster, R., Sutcliffe, M., Carswell, D., Hauser, C., & Barras, J. (2020). An improved methodology of melt pool monitoring of direct energy deposition processes. *Optics and Laser Technology*, *127*(December 2019).

[26]. Lane BM, Yeung H (2020) Process Monitoring Dataset from the Additive Manufacturing Metrology Testbed (AMMT): “Overhang Part X4.” *NIST Journal of Research (JRES)*. https://doi.org/(publication pending

[27]. NIST AMMT Datasets, https://www.nist.gov/el/ammt-temps/datasets

[28]. ISO/ASTM 52921, "Standard Terminology for Additive Manufacturing—Coordinate Systems and Test Methodologies", https://compass.astm.org/EDIT/html_annot.cgi?ISOASTM52921+13 (2019)

Measurements of the Most Significant Software Security Weaknesses

Carlos Cardoso Galhardo
National Institute of Standards and Technology; INMETRO
carlos.cardosogalhardo@nist.gov;cegalhardo@inmetro.gov.br

Peter Mell
National Institute of Standards and Technology
peter.mell@nist.gov

Irena Bojanova
National Institute of Standards and Technology
irena.bojanova@nist.gov

Assane Gueye
UADB-Senegal & Prometheus Computing
assane1.gueye@uadb.edu.sn

ABSTRACT

In this work, we provide a metric to calculate the most significant software security weaknesses as defined by an aggregate metric of the frequency, exploitability, and impact of related vulnerabilities. The Common Weakness Enumeration (CWE) is a well-known and used list of software security weaknesses. The CWE community publishes such an aggregate metric to calculate the 'Most Dangerous Software Errors'. However, we find that the published equation highly biases frequency and almost ignores exploitability and impact in generating top lists of varying sizes. This is due to the differences in the distributions of the component metric values. To mitigate this, we linearize the frequency distribution using a double log function. We then propose a variety of other improvements, provide top lists of the most significant CWEs for 2019, provide an analysis of the identified software security weaknesses, and compare them against previously published top lists.

CCS CONCEPTS

• **Security and privacy** → **Vulnerability management**; **Software and application security**.

KEYWORDS

Security, Weakness, Software Flaw, Severity

ACM Reference Format:
Carlos Cardoso Galhardo, Peter Mell, Irena Bojanova, and Assane Gueye. 2020. Measurements of the Most Significant Software Security Weaknesses . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3427228.3427257

1 INTRODUCTION

In 2019, there were over 17 000 documented software vulnerabilities [22] that enable malicious activity. While many are discovered, they map to a relatively small set of underlying weakness types. We posit that if the most significant of these types can be identified, developers of programming languages, software, and security tools can focus on preventing them and thus over time diminish the quantity and severity of newly discovered vulnerabilities.

In this work, we provide a metric to calculate the most significant security weaknesses (MSSW) in software systems. We define a 'significant' weakness as one that is both frequently occurring among the set of publicly published vulnerabilities and results in high severity vulnerabilities (those that are easily exploitable and have high impact). The set of security weakness types upon which we calculate significance comes from the Common Weakness Enumeration (CWE) [15]. We also leverage the Common Vulnerabilities and Exposures (CVE) [13] repository of publicly announced vulnerabilities, the Common Vulnerability Scoring System (CVSS) [7] to measure the severity of vulnerabilities, and the National Vulnerability Database (NVD) [22] to map the CVEs to both CWEs and CVSS scores.

In the fall of 2019, the CWE community published an equation to calculate the 'Top 25 Most Dangerous Software Errors' (MDSE) among the set of CWEs [17]. It follows the form of the common security risk matrix combining probability and severity (e.g., [5]). The MDSE equation claims to combine 'the frequency that a CWE is the root cause of a vulnerability with the projected severity'; the equation description implies that both factors are weighed equally (making no mention of any bias). However, we empirically find that the equation highly biases frequency and almost ignores severity in generating top lists of varying sizes. This is due to the equation multiplying calculated frequency and severity values together though each has has very different distributions. Frequency distributions have a power law like curve, while severity distributions are more uniform. Our mitigation is to create a revised equation, named MSSW, that adjusts the frequency distribution using a double log function to better match it to the severity distribution. We also fix an error in how normalization is done in the MDSE equation.

We next improve upon the data collection approach used by the MDSE equation by leveraging published literature [12]. Lastly, we publish top lists of the most significant CWEs for 2019, provide an analysis of those software security weaknesses, and compare our top lists against previously published lists. It is our hope that our data and methodology will be adopted to focus our collective security resources in reducing the most significant software security weaknesses.

The rest of this work is organized as follows. Section 2 provides background on CVE, CVSS, CWE, NVD, and the MDSE equation. Section 3 discusses the limitations of the MDSE equation. Section 4 presents our MSSW equation that mitigates the previously identified limitations. Section 5 provides two lists of the most significant CWEs at two different levels of software flaw type abstractions. Section 6 provides a discussion and analysis of the most significant CWEs identified. Section 7 presents related work, Section 8 discussed possible future research, and Section 9 concludes.

Cardoso Galhardo, Carlos; Mell, Peter; Bojanova, Irena; Gueye, Assane. "Measurements of the Most Significant Software Security Weaknesses."
Presented at Annual Computer Security Applications Conference (ACSAC 2020). December 07, 2020 - December 11, 2020.

2 BACKGROUND

2.1 Common Vulnerabilities and Exposures

The CVEs are a large set of publicly disclosed vulnerabilities in widely-used software. They are enumerated with a unique identifier, described, and referenced with external advisories [13] [1].

2.2 Common Vulnerability Scoring System

CVSS 'provides a way to capture the principal characteristics of a vulnerability and produce a numerical score reflecting its severity' [6]. The CVSS base score reflects the inherent risk of a vulnerability apart from any specific environment. The base score is composed from two sub-scores that calculate exploitability (how easy it is to use the vulnerability in an attack) and impact (how much damage the vulnerability can cause to an affected component).

The exploitability score is determined by the following:

- attack vector: 'the context by which vulnerability exploitation is possible',
- attack complexity: 'the conditions beyond the attacker's control that must exist in order to exploit the vulnerability',
- privileges required: 'the level of privileges an attacker must possess before successfully exploiting the vulnerability', and
- user interaction: a human victim must participate for the vulnerability to be exploited.

The impact score is determined by measuring the impact to the confidentiality, integrity, and availability of the affected system. Also included is a scope metric that 'captures whether a vulnerability in one vulnerable component impacts resources in components beyond its security scope'. The specifics on these metrics and the details for the three equations can be found in the CVSS version 3.1 specification at [7].

2.3 Common Weakness Enumeration

The Common Weakness Enumeration (CWE) [10] is 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' [15]. It contains an enumeration, descriptions, and references for 839 software weaknesses that are referred to as CWEs, where each is labelled CWE-X with X being an integer.

The CWE weaknesses model has four layers of abstraction: pillar, class, base, and variant. There is also the notion of a compound, that associates two or more interacting or co-occurring CWEs [18]. These abstractions reflect to what extent issues are described in terms of five dimensions: behavior, property, technology, language, and resource. Variant weaknesses are at the most specific level of abstraction; they describe at least three dimensions. Base weaknesses are more abstract than variants and more specific than classes; they describe two to three dimensions. Class weaknesses are very abstract; they describe one to two dimensions, typically not specific about any language or technology. Pillar weaknesses are the highest level of abstraction.

There are a set of taxonomies, called views, to help organize the CWEs. Two prominent CWE taxonomies are the 'Research Concepts' (view 1000) and 'Development Concepts' (view 699). There is also a view 1003 that was made specifically to describe the set of CVEs that contains 124 CWEs. It is called 'CWE Weaknesses for Simplified Mapping of Published Vulnerabilities View'.

2.4 National Vulnerability Database

The CWE effort uses the National Vulnerability Database (NVD) [22] as a repository of data from which to calculate the MDSE scores. The NVD contains all CVEs and for each CVE it provides a CVSS score along with the applicable CWE(s) that describe the weakness(es) enabling the vulnerability. For the empirical work in this paper, we use the complete set of 17 308 CVEs published by NVD for 2019, that were available as of 2020-03-19.

2.5 Most Dangerous Software Error Equation

The MDSE equation is designed to balance the frequency and severity in ranking the CWEs. The frequency is determined by the number of CVEs that map to a given CWE in the time period of study. The severity is determined by the mean CVSS score for the CVEs mapped to a given CWE. The MDSE score for a CWE is produced by multiplying the normalized frequency by the normalized severity and then multiplying by 100. We now describe this metric more formally.

2.5.1 *Metric for Normalized Frequency.*

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, defined as follows:

$$N_i = \sum_{j \in J} e_{ij}, \tag{1}$$

where

$$e_{ij} = \begin{cases} 1, & \text{if CVE } j \text{ is mapped to CWE } i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Now let F_i be the normalized frequency for CWE i, defined as follows:

$$F_i = \frac{N_i - \min_{i' \in I}(N_{i'})}{\max_{i' \in I}(N_{i'}) - \min_{i' \in I}(N_{i'})}. \tag{3}$$

2.5.2 *Metric for Normalized Severity.*

Let J, N_i, and e_{ij} be as defined above in Section 2.5.1. Let s_j be the CVSS base score for CVE j. For CWE $i \in I$, let $\overline{S_i}$ be the mean CVSS score, defined as follows:

$$\overline{S_i} = \frac{\sum_{j \in J} s_j e_{ij}}{N_i}. \tag{4}$$

Now let S_i be the normalized severity for CWE i, defined as follows:

$$S_i = \frac{\overline{S_i} - \min_{j \in J}(s_j)}{\max_{j \in J}(s_j) - \min_{j \in J}(s_j)}. \tag{5}$$

2.5.3 *Most Dangerous Software Error Metric.*

Let $MDSE_i$ be the MDSE score for CWE i, defined as follows:

$$MDSE_i = F_i * S_i * 100. \tag{6}$$

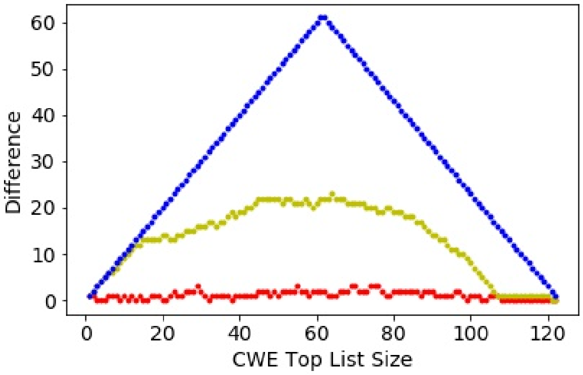


Figure 1: The Size of the Set Difference between Top Lists from the MDSE Equation Compared to Frequency Top Lists (red bottom line), Severity Top Lists (yellow middle line), and the Theoretical Maximum (blue top line)

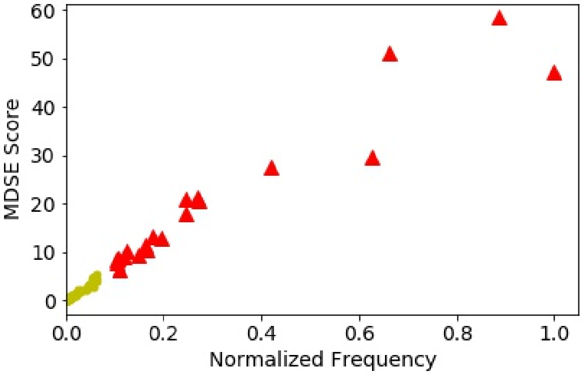


Figure 2: CWEs Chosen (Red Triangles) and Not Chosen (Yellow Circles) for a MDSE Top 20 List Relative to Frequency

3 LIMITATIONS OF THE EQUATION

The MDSE equation was designed to and appears to combine both frequency and severity in determining the individual scores used to rank the CWEs. The frequency component is calculated in equation 3 and the severity component is calculated in equation 5; both are brought together in equal proportions in equation 6 to create the MDSE score. And both the severity and frequency are normalized in equations 3 and 5 to ensure that their scales match for the multiplication in equation 6.

However, we empirically find that the MDSE equation strongly biases frequency over severity. To demonstrate this, we calculate MDSE top CWE lists for all possible list sizes. While there exist 839 CWEs, the CVE data used as MDSE input is mapped only to 124 view 1003 CWEs (see section 2.3)[1]. Thus the maximum top list size is 124. We also calculate top CWE lists using just the frequency equation 3 and then just the severity equation 5. For each CWE top list size, we perform a set difference between the MDSE top list and the frequency top list. We then also do this between the MDSE top list and the severity top list. The size of the set difference between the MDSE top list and the frequency top list (for all possible top list sizes) has a maximum difference of 3. The size of the set difference between the MDSE top list and the severity top list (for all possible top list sizes) has a maximum difference of 23. This is shown graphically in Figure 1. The bottom red line represents the set difference using frequency and the yellow middle line represents the set difference using severity. The top blue line shows the maximal possible set difference that could be achieved using the 124 CWEs.

More qualitatively, the red line hovers close to a y-axis value of 0 which means that for all list sizes the top list generated using just frequency is almost identical to the top list generated using the MDSE equation. The middle yellow line being far from the y-axis value of 0 means that for all list sizes the top list generated using just severity is very different from the top list generated using the MDSE equation. Note that the yellow line shows an almost maximal difference for top list sizes of up to 15.

3.1 Limitation 1: Distribution Differences

The MDSE equation in practice biases frequency over severity, even though its equations treat them equally, because frequency and severity have very different distributions. The frequency distribution has the majority of CWEs at a very low frequency and a few at a very high frequency (somewhat resembling a power law curve). This can be seen in Figure 2 by looking at how each CWE maps to the x-axis (note that most of the yellow dots overlap, there are 102 yellow dots and 20 red triangles). The figure shows the MDSE scores for each CWE and shows how (for a top list of size 20) the top scoring chosen CWEs are exactly the most frequent CWEs. This is not unique and occurs for many top lists (e.g., for sizes 11, 13, 15, 16, 20, 21, 32, and 38) as shown when the bottom red line is at 0 in Figure 1. The other sizes of top lists produce graphs that are almost identical to that in Figure 2, with at most 3 yellow circles just to the right of the leftmost red triangles representing the chosen CWEs.

The severity distribution is more uniform within a limited range. It can be seen in Figure 3 by looking at how the CWEs map to the x-axis. This figure shows how the top MDSE scoring chosen CWEs do not necessarily map to the CWEs with the highest severity. In fact, only 1 of the top 10 most severe CWEs made the MDSE top 20 list (note that many of the yellow circles lay on top of each other).

3.2 Limitation 2: Normalization Error

Equation 5 normalizes S_i based on the maximum and minimum CVSS score found in the set of inputted CVEs. However, this does not lead to the expected and desired normalized distribution from 0 to 1. For our data the range is from .28 to .97, as can be seen from the mappings of the points onto the x-axis in Figure 3. The reason for this is that $\overline{S_i}$ has a smaller range than the maximum and minimum CVSS score because each $\overline{S_i}$ represents the mean of the CVSS score for the CVEs that map to CWE i. This limitation,

[1] This is expected as view 1003 was designed to cover the types of vulnerabilities in CVE.

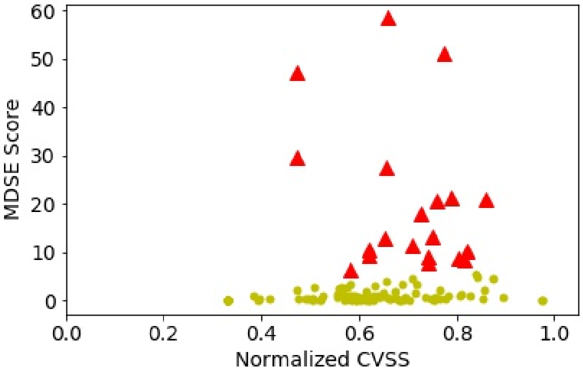


Figure 3: CWEs Chosen (Red Triangles) and Not Chosen (Yellow Circles) for a MDSE Top 20 List Relative to Severity

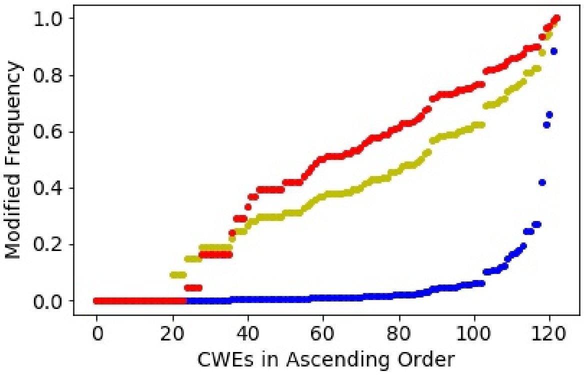


Figure 4: Normalized Distributions of Frequency (bottom blue line), Log of Frequency (middle yellow line), and Double Log of Frequency (top red line).

while of less consequence than the previous, constrains the range of S_i values thus further lessening the influence that severity has in determining a MDSE score.

4 MITIGATED EQUATION

We mitigate the limitations of the MDSE equation by replacing equations 3, 5, and 6 with the five equations that follow:

$$k = \frac{1}{\log_e \log_e \max\limits_{i \in I}(N_i)}, \tag{7}$$

$$F'_i = \begin{cases} \log_e N_i, & \text{if } N_i >= 1, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

$$F''_i = \begin{cases} k \log_e F'_i, & \text{if } F'_i >= 1, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

$$S'_i = \frac{\overline{S_i} - \min\limits_{i' \in I}(\overline{S_{i'}})}{\max\limits_{i' \in I}(\overline{S_{i'}}) - \min\limits_{i' \in I}(\overline{S_{i'}})}, \tag{10}$$

$$MSSW_i = F''_i * S'_i * 100. \tag{11}$$

4.1 Explanation of Mitigated Equation

Equation 8 takes the log of the frequency using the natural log as the base. Equation 9 then takes the log of equation 8, again using the natural log as a base and multiplies the result by k (from equation 7). The k coefficient serves the purpose of normalizing the resulting values between 0 and 1 (to match the severity range in equation 10).

These three equations modify the power law like frequency distribution to make it more linear, thus addressing limitation 1 (from Section 3.1). This can be seen in Figure 4. Each value on the x-axis represents a particular CWE, ordered from least frequent to most frequent. The lower blue line represents the normalized frequency (i.e., number of CVEs mapped to a particular CWE). Note the slow increase in frequency up to the 100th CWE, followed by a rapid increase terminating in an almost vertical line (i.e. large derivative). This behavior creates large differences between the most frequent CWEs and almost no difference between the lowest CWEs. If $g(x) = \log f(x)$ its derivative is

$$\frac{dg(x)}{dx} = \frac{1}{f(x)} \frac{df(x)}{dx}, \tag{12}$$

thus applying a log function over the frequency should minimize differences between the most frequent CWEs.

The middle yellow line represents taking the log of the frequency (equations not shown), which helps linearize but still results in an upwards curve on the right side. Thus, we apply a double log for further linearization (see the top red line). We note that this approach is not pseudo-linear for the most infrequent of CWEs. However, this does not cause problems as our goal is to identify the most significant and any such CWE must have at least a moderate frequency.

Our modified MDSE equation 11 then multiplies frequency and severity as in the original MDSE equation, but it multiplies from two distributions that have a similar shape for the part of the functions that are of interest. This enables the MSSW equation to more fairly balance evaluating frequency and severity in scoring and ranking a CWE.

To address limitation 2 from Section 3.2, equation 10 normalizes the severity using the maximum and minimum mean severity values. This gives the distribution a full 0 to 1 range which is not achieved in the MDSE equation 5.

Equation 11 is our final modified MDSE equation. We recommend its use in place of the published MDSE equation.

4.2 Analysis of Mitigated Equation

We now conduct three experiments to evaluate the effect of the MSSW equation in making the frequency and severity distributions more similar and in producing top lists with more equal inclusion of both frequency and severity. A fourth experiment involving

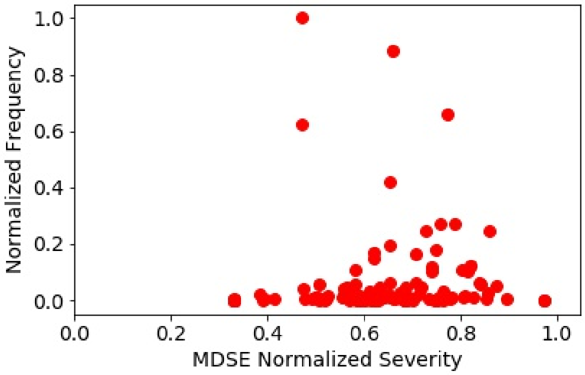


Figure 5: MDSE Equation Risk Map

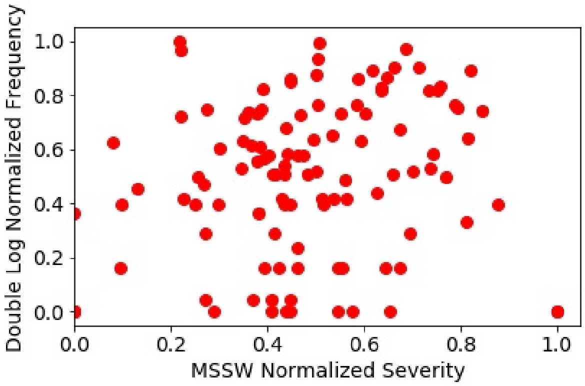


Figure 6: MSSW Equation Risk Map

correlation calculations is provided in Section 5 (because it includes some variants introduced in that section).

4.2.1 Risk Map Experiment. Figure 5 shows an MDSE risk map for the evaluated CWEs. Each red dot represents a CWE positioned according to its S_i severity and F_i frequency. In general, CWEs towards the upper right are more significant and those towards the lower left are less significant. Note how the majority of the CWEs are squished very close to the x-axis as many have a very small frequency. Also, the range of x-values is constrained from .37 to .97 (when the normalization should make it from 0 to 1).

Figure 6 shows the same risk map using our double log frequency F_i'' and our modified severity S_i'. Note how the CWEs are now more uniformly spread over the y-axis. Also, the range of x-axis values is now from 0 to 1. The MSSW equation that combines frequency and severity using the values shown in Figure 6 will now more equally combine them than with the MDSE values shown in Figure 5.

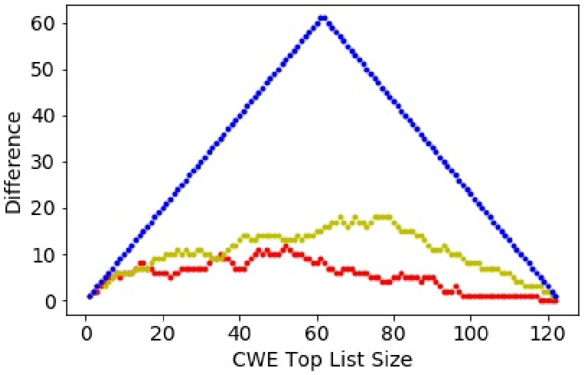


Figure 7: The Size of the Set Difference between Top Lists from the MSSW Equation Compared to Frequency Top Lists (red lower line), Severity Top Lists (yellow middle line), and the Theoretical Maximum (blue top line)

4.2.2 Set Difference Experiment. In Figure 7 we show the size of the set difference between the MSSW top list and the severity top list (the mostly lower red line). We also calculate the set difference between the MSSW top list and the frequency top list (the middle yellow line). Note how the red and yellow lines are much closer together than in Figure 1 and how the red line does not hover close to 0 like it does in Figure 1. This demonstrates that the MSSW equation is more evenly balancing inclusion of the top frequency and top severity CWEs.

Note that the goal is not to have the red and yellow lines match. The top list should not necessarily evenly include an equal number of both top frequency and top severity CWEs. Our point with this analysis is to show how the MDSE equation almost exclusively chooses the top frequency CWEs and how our MSSW equation factors in CWEs from both sets. The next subsection will evaluate this more equal inclusion in more detail, focusing on top lists of size 20.

4.2.3 Chosen CWE Experiment. Figure 8 shows the MSSW scores plotted against the double log frequency F_i'' scores. Each point represents a CWE. The red triangles indicate the CWEs that were chosen for the MSSW top 20 list. Note how unlike in the analogous Figure 2 for MDSE, there are many higher frequency CWEs that are not chosen for the top 20 list due to their severity not being high enough.

Likewise, Figure 9 shows the MSSW scores plotted against the S_i' normalized mean CVSS score for each CWE. Note how the range spreads from 0 to 1, unlike the analogous Figure 3 for the MDSE equation. Also note how the MSSW equation chooses CWEs for the top 20 list from CWEs with generally higher CVSS scores. However, it excludes many high severity CWEs because their frequencies were too low.

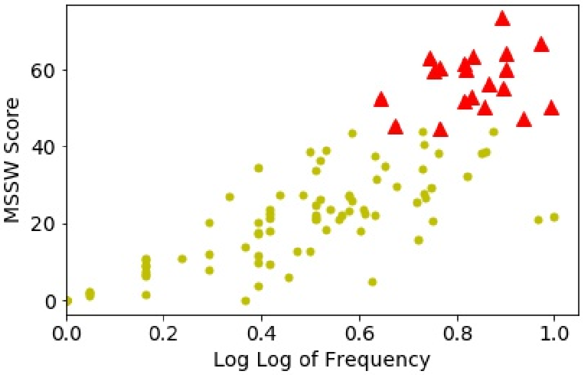


Figure 8: CWEs Chosen (Red Triangles) and Not Chosen (Yellow Circles) for a MSSW Top 20 List Relative to Frequency

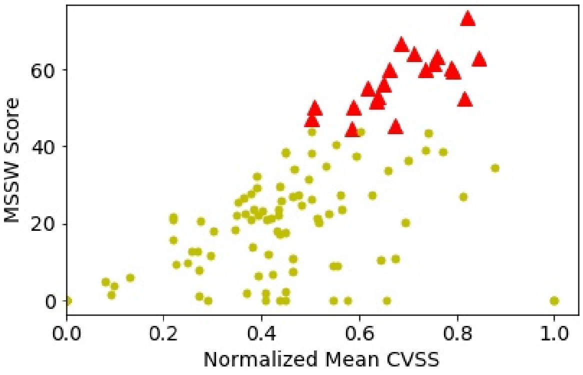


Figure 9: CWEs Chosen (Red Triangles) and Not Chosen (Yellow Circles) for a MSSW Top 20 List Relative to Severity

5 2019 TOP 20 LISTS OF THE MOST SIGNIFICANT WEAKNESSES

We now use our MSSW equation to generate lists of the most significant software security weaknesses. We choose a list size of 20, somewhat arbitrarily, to enable the lists to conveniently fit on a page. We performed the experiments on a variety of list sizes and did not discover any appreciable differences. We did not choose a size of 25 to match the CWE top list because we needed to produce two top lists of differing levels of abstraction to get the most accurate results (explained below) and thus were unable to produce a single list of 25 CWEs for an ideal comparison with the CWE top 25 list.

We follow the approach in [12] of separately providing a top list for CWEs of higher levels of abstraction (pillars and classes) apart from a list covering CWEs of lower levels of abstraction (bases, variants, and compounds). This was done in [12] to avoid errors in frequency calculations that exist in CWE's top 25 list. The paper argues that CWEs mapped to lower level abstractions (e.g., bases) also should count towards their parent abstractions (e.g., classes); this is not done with MDSE calculations. For example, class CWE-20 (Improper Input Validation) is a parent of base CWE-1289 (Improper Validation of Unsafe Equivalence in Input). If a vulnerability exists with CWE-1289, then CWE-20 also needs to be taken into account with the CWE frequency counts. However when this frequency propagation is performed, combining together the two abstractions results in a single top list with a bias towards parents with many children (especially popular children). Thus the 2 levels of abstraction need to be presented in separate top lists.

We will refer to the higher level abstraction list as the class list and the lower level abstraction list as the base list for convenience and because both lists are primarily composed of either classes or bases. We also follow [12] in using published CWE taxonomy views 1000 and 1008 (discussed in Section 2.3) to propagate CVE data from child CWEs to their parents (discussed above). This provides a more accurate mapping of CVEs onto the CWEs, providing a more accurate data foundation upon which to apply our MSSW equation.[2]

These modifications also alter the frequency and severity distributions which could potentially render our double log function invalid. However, Table 1 shows correlation results for using and not using all combinations of the modifications adopted from [12]. It shows that the MDSE equation is highly correlated to frequency (.97 or higher) with very little correlation to severity (.25 or lower) regardless of the modifications used or not used. It also shows that the MSSW equation is strongly correlated to both frequency (.81 or higher with one exception) and severity (.66 or higher) regardless of the modifications used. Our one exception is for the class list using propagation with MSSW; even here the frequency correlation was .55 (still strong but much less than the others).

Note that our objective is not for the correlations to necessarily be equal, but that there exists a strong correlation for both frequency and severity. Depending upon the data, the higher frequency CVEs may or may not also be the highest severity CVEs. If so, then the correlations to frequency and severity would both be very high and almost equal. If not, both should still be high but one may be higher than the other. What we do not want in these results is for one of frequency or severity to have a high correlation and the other to have a very low correlation (which can be seen with the MDSE equation).

We also checked to see that the double log still linearized the frequency distribution when using both variants from [12]. While propagating CVEs over the CWEs using the CWE taxonomies and using all applicable CWEs (i.e., pillars, classes, bases, variants, and compounds), the results show that the double log does still linearize the frequency (see Figure 10). The same results were obtained while also performing the experiment using just the pillars/classes and then just the bases, variants, and compounds (graphs not shown).

[2] This propagation especially helps the formulation of the class list since most classes have children. It has a lesser effect on the base list. Note that it is impossible to inverse the propagation of data. CWEs are labelled as specifically as possible by NVD analysts so CVEs described by pillar or class CWEs do not get reflected in the base list. It is even possible that they shouldn't because there may be unidentified bases missing from view 1003 that are still covered by the view 1003 classes.

Table 1: Measurements Showing the Pearson Correlation of MDSE and MSSW to Frequency and Severity

			Correlation	
Equation	Abstraction	Propagation	Frequency	Severity
MDSE	All	Yes	.99	.08
MDSE	All	No	.98	.18
MDSE	High	Yes	.99	.10
MDSE	High	No	.98	.25
MDSE	Low	Yes	.97	.20
MDSE	Low	No	.97	.18
MSSW	All	Yes	.81	.70
MSSW	All	No	.86	.66
MSSW	High	Yes	.55	.96
MSSW	High	No	.84	.68
MSSW	Low	Yes	.84	.67
MSSW	Low	No	.83	.68

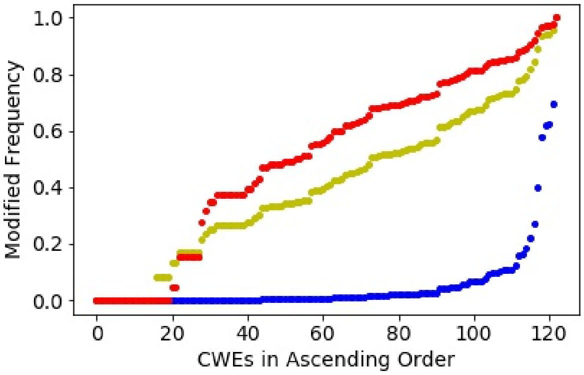


Figure 10: Normalized Distributions of Frequency (bottom blue line), Log of Frequency (middle yellow line), and Double Log of Frequency (top red line).

Using our MSSW equation to aggregate the frequency and severity of CWEs, the top 20 class list for 2019 is shown in Table 2. The top 20 base list is shown in Table 3. These two lists use the modification from [12] where the CVEs are propagated up through the CWE taxonomies. We claim that these two lists represent the most accurate measurement yet produced for determining the most significant software security weaknesses. Given that there is no ground truth for how to best combine frequency and severity and no ground truth upon which to establish the correctness of the CVSS metric, it is likely impossible to prove any such metric as maximally effective. We make our 'most accurate measurement yet' claim based on the demonstrated limitations in the published MDSE equation and a lack of competing published alternatives.

6 DISCUSSION AND ANALYSIS OF THE MOST SIGNIFICANT WEAKNESSES

In this section, we evaluate our 2019 MSSW class and base lists (see Tables 2 & 3) and compare them against the 2019 CWE Top 25 MDSE List [17] (reproduced in Table 4).

As stated previously, we expect the MDSE list to vary from the MSSW class and base lists because:

(1) the MDSE list is biased towards the frequency of a CWE occurring in CVEs,
(2) we use the taxonomy propagation approach from [12], and
(3) the class and base lists contain a total of 40 CWEs while the MDSE list contains 25 CWEs.

6.1 High Level Summaries

View 1003 contains two pillars (CWE-682 and CWE-697) and 36 classes, as well as 81 bases, three variants (CWE-415, CWE-416, and CWE-401), and two compounds (CWE-352 and CWE-384).

The MDSE Top 25 [17] ranks CWE items across all the layers of abstraction from view CWE-1003. The list has seven classes, 16 bases, one variant, and one compound. Interestingly, some of these top CWEs have child-parent relationships among themselves.

A simple inspection of the list shows how parent CWEs do not receive CVE counts from their children. For example, the count for the top class CWE-119 (rank 1, count 1048) does not include the counts of its children CWE-125 (rank 5, count 678) and CWE-787 (rank 12, count 473). Analogously, the count for the class CWE-287 (rank 13, count 299) does not include the counts of its children base CWE-798 (rank 19, count 91) and base CWE-295 (rank 25, count 77).

Our MSSW class list is comprised of 19 class CWEs and the pillar CWE-682 (rank 9) – see Table 2. Only three CVEs are directly described with the pillar, but it appears in the list because there is a set of severe CVEs described with its children (see subsection 6.5). Our MSSW base list is comprised of 17 bases, the variants CWE-416 (rank 7) and CWE-415 (rank 14), and the compound CWE-352 (rank 10) – see Table 3. Each of the two lists properly compare items of the same kind. Interestingly but not surprisingly, each CWE from the base list is a child of a CWE from class list. However, the ordering of these parent-child pairs are not necessarily preserved between the two lists.

6.2 Set Differences

There are differences in the set of CWEs covered by our top 20 MSSW class and base lists and the MDSE list. The pillars/classes in the MDSE list that do not appear in the class list are: CWE-200 and CWE-732. The bases/variants/compounds in the MDSE list that do not appear in the base list are: CWE-79, CWE-476, CWE-772, CWE-426, and CWE-295. The base list contains base CWE-120 (a child of CWE-119), which does not appear in the MDSE list.

Note that the two classes from the MDSE list with children in the same list are also in the class list (emphasizing their importance): class CWE-119 with children base CWE-125 and base CWE-787, and class CWE-287 with children base CWE-798 and base CWE-295.

Table 2: 2019 MSSW Top 20 Pillars/Classes, Propagating CVSS Data over CWE Taxonomies

Rank	Identifier	CWE Description	MSSW Score	Frequency	Mean CVSS
1	CWE-913	Improper Control of Dynamically-Managed Code Resources	78.31	188	8.81
2	CWE-119	Improper Restriction of Operations within Bounds of a Memory Buffer	71.14	2745	8.00
3	CWE-669	Incorrect Resource Transfer Between Spheres	64.86	181	8.31
4	CWE-672	Operation on a Resource after Expiration or Release	64.56	876	7.96
5	CWE-330	Use of Insufficiently Random Values	63.74	111	8.43
6	CWE-704	Incorrect Type Conversion or Cast	62.55	54	8.68
7	CWE-287	Improper Authentication	59.75	627	7.86
8	CWE-345	Insufficient Verification of Data Authenticity	54.60	483	7.73
9	CWE-682	Incorrect Calculation	51.94	215	7.78
10	CWE-269	Improper Privilege Management	50.57	258	7.70
11	CWE-610	Externally Controlled Reference to a Resource in Another Sphere	48.38	725	7.46
12	CWE-706	Use of Incorrectly-Resolved Name or Reference	39.04	358	7.23
13	CWE-20	Improper Input Validation	38.56	3960	6.99
14	CWE-116	Improper Encoding or Escaping of Output	32.13	2461	6.82
15	CWE-400	Uncontrolled Resource Consumption	32.07	272	7.01
16	CWE-74	Improper Neutralization of Special Elements in Output ... ('Injection')	32.06	2455	6.82
17	CWE-754	Improper Check for Unusual or Exceptional Conditions	32.05	264	7.01
18	CWE-326	Inadequate Encryption Strength	28.21	35	7.24
19	CWE-668	Exposure of Resource to Wrong Sphere	26.59	2292	6.66
20	CWE-436	Interpretation Conflict	22.40	17	7.19

Table 3: 2019 MSSW Top 20 Bases/Variants/Compounds, Propagating CVSS Data over CWE Taxonomies

Rank	Identifier	CWE Description	MSSW Score	Frequency	Mean CVSS
1	CWE-89	Improper Neutralization of Special Elements used ... ('SQL Injection')	71.70	384	8.89
2	CWE-502	Deserialization of Untrusted Data	61.73	83	9.01
3	CWE-787	Out-of-bounds Write	61.57	423	8.34
4	CWE-78	Improper Neutralization of Special ... ('OS Command Injection')	61.22	194	8.58
5	CWE-120	Buffer Copy without Checking Size of ... ('Classic Buffer Overflow')	59.35	162	8.55
6	CWE-94	Improper Control of Generation of Code ('Code Injection')	58.62	100	8.72
7	CWE-798	Use of Hard-coded Credentials	58.07	89	8.75
8	CWE-434	Unrestricted Upload of File with Dangerous Type	57.95	167	8.46
9	CWE-416	Use After Free	56.69	426	8.09
10	CWE-352	Cross-Site Request Forgery (CSRF)	51.60	386	7.86
11	CWE-346	Origin Validation Error	51.51	430	7.82
12	CWE-613	Insufficient Session Expiration	51.08	402	7.82
13	CWE-190	Integer Overflow or Wraparound	48.79	164	7.95
14	CWE-415	Double Free	43.17	46	8.15
15	CWE-125	Out-of-bounds Read	42.34	658	7.28
16	CWE-129	Improper Validation of Array Index	41.97	25	8.50
17	CWE-611	Improper Restriction of XML External Entity Reference	41.47	100	7.69
18	CWE-918	Server-Side Request Forgery (SSRF)	41.05	74	7.78
19	CWE-22	Improper Limitation of a Pathname to a Restricted ... ('Path Traversal')	39.40	309	7.27
20	CWE-191	Integer Underflow (Wrap or Wraparound)	37.76	18	8.47

6.3 Reordered Rankings

The relative orderings in the MDSE list often do not match the orderings in the MSSW class and base lists. There are some notable reorderings. CWE-89 (Structured Query Language (SQL) Injection) and CWE-502 (Deserialization of Untrusted Data) climb up in the base list due to their highest severities of 8.89 and 9.01. CWE-913 (Improper Control of Dynamically-Managed Code Resources) does not even appear in the MDSE Top 25 list, as it has only three direct occurrences in the CVEs. However, it climbs up to first position in the class list due to its highest severity of 8.81 and its 188 propagated occurrences. Its main child contributor is base CWE-502 with frequency of 83 and severity of 9.01. CWE-119 (Improper Restriction of Operations within the Bounds of a Memory Buffer) in the MDSE list, while widely used with 2745 propagated occurrences in the CVEs, is quite less severe than CWE-913 and drops to rank 2 in the MSSW class list.

Table 4: Reproduction of the 2019 CWE Top 25 Most Dangerous Software Errors List[17]

Rank	Identifier	CWE Description	MDSE Score
1	CWE-119	Improper Restriction of Operations within the Bounds of a Memory Buffer	75.56
2	CWE-79	Improper Neutralization of Input During Web Page Generation ('Cross-site Scripting')	45.69
3	CWE-20	Improper Input Validation	43.61
4	CWE-200	Information Exposure	32.12
5	CWE-125	Out-of-bounds Read	26.53
6	CWE-89	Improper Neutralization of Special Elements used in an SQL Command ('SQL Injection')	24.54
7	CWE-416	Use After Free	17.94
8	CWE-190	Integer Overflow or Wraparound	17.35
9	CWE-352	Cross-Site Request Forgery (CSRF)	15.54
10	CWE-22	Improper Limitation of a Pathname to a Restricted Directory ('Path Traversal')	14.1
11	CWE-78	Improper Neutralization of Special Elements used in an OS Command ('OS Command Injection')	11.47
12	CWE-787	Out-of-bounds Write	11.08
13	CWE-287	Improper Authentication	10.78
14	CWE-476	NULL Pointer Dereference	9.74
15	CWE-732	Incorrect Permission Assignment for Critical Resource	6.33
16	CWE-434	Unrestricted Upload of File with Dangerous Type	5.5
17	CWE-611	Improper Restriction of XML External Entity Reference	5.48
18	CWE-94	Improper Control of Generation of Code ('Code Injection')	5.36
19	CWE-798	Use of Hard-coded Credentials	5.12
20	CWE-400	Uncontrolled Resource Consumption	5.04
21	CWE-772	Missing Release of Resource after Effective Lifetime	5.04
22	CWE-426	Untrusted Search Path	4.4
23	CWE-502	Deserialization of Untrusted Data	4.3
24	CWE-269	Improper Privilege Management	4.23
25	CWE-295	Improper Certificate Validation	4.06

6.4 The Two Most Dangerous CWEs: Injection vs. Memory Errors

The two most distinctive groups of weaknesses both in the MDSE Top 25 list and the two MSSW Top 20 lists are injection and memory errors. However, the use of the MSSW equation and the split into class and base lists considerably reorders these two groups, as well as brings in new CWEs and drops some CWEs.

6.4.1 Injection Weaknesses. Injection is the most dangerous type of weakness, represented by bases) CWE-89 (SQL Injection), CWE-502 (Deserialization of Untrusted Data), CWE-78 (OS Command Injection), CWE-94 (Code Injection), and CWE-611 (Improper Restriction of Extensible Markup Language (XML) External Entity Reference), with ranks 1, 2, 4, 6, and 17 respectively in the base list (see Table 3). The MDSE list also contains these five CWEs, however the rankings of the first three are 6, 16, and 11 due to their lower frequencies of 397, 85, and 217. The MSSW inclusion of their high severity scores of 8.89, 9.01, and 8.58 moved them several positions up in the base list. Note that CWE-502 covers Object Injection.

Also of importance is that the second ranked in the MDSE list CWE-79 (Cross-site Scripting), is not in our MSSW base list. Although it has the highest frequency of 1571, its severity score of 5.83 is relatively low.

The MSSW class list includes CWE-913 (Improper Control of Dynamically-Managed Code Resources), CWE-116 (Improper Encoding or Escaping of Output), and CWE-74 (Injection), ranked 1, 14, and 16 (see Table 2). The reason for that is CWE-913 is the parent of CWE-502, CWE-116 is a typical cause of injection and CWE-74 is the parent of CWE-78, CWE-89, and CWE-94. Interestingly, the class CWE-74 has rank 16 among classes, while its children bases CWE-89, CWE-78, and CWE-94 are ranked 1, 4, 6 among bases. The frequencies of 2455 for CWE-74, 384 for CWE-89, 194 for CWE-78, and 100 for CWE-94, leave 1777 injection CVEs that are described with CWEs that are either very infrequent or not severe. These are bases CWE-79 (Cross-site Scripting) with the low severity of 5.83, CWE-88 (Argument Injection) with the low frequency of 6, and CWE-91 (XML Injection) with the low frequency of 16. Being not too dangerous they bring the class CWE-74 down to rank 16. That same base CWE-79, not included in the MSSW base list, is ranked 2nd in the MDSE list due to the frequency biasing.

6.4.2 Memory Weaknesses. The most dangerous memory weaknesses are CWE-787 (Out-of-bounds Write) and CWE-120 (Classic Buffer Overflow) with ranks 3 and 5 – see Table 3. Both of them are included in the base list but not the MDSE list, due to the correction of the frequency bias towards proper inclusion of their severity scores of 8.34 and 8.55.

The other memory weaknesses in the MSSW class and base lists are as follows:

- bases CWE-125 and CWE-787 are buffer overflow (out of bounds read or write)
- variant CWE-416 is use after free (use of deallocated memory through a dangling pointer)

- variant CWE-415 is double free (deallocate of already deallocated memory)
- class CWE-119 is a general memory corruption weakness, which includes buffer overflow, use after free and double free.
- class CWE-400 is memory overflow (stack/heap exhaustion) [21]

6.4.3 Injection/Memory Weakness Comparison. Compared to MDSE, the MSSW equation brings up several injection weaknesses with much higher severity than that of any memory weaknesses. The related CVE analysis confirms that the injection CVE are easier to exploit and with higher impact. An injection directly leads to arbitrary command, code, or script execution. Once a SQL injection is in place, there is no need of additional sophisticated attack crafting or use of glitches in the system. However, it takes considerable extra effort for an attacker to turn a buffer overflow into an arbitrary code execution. He or she would need to have exceptional skills, such as to apply spraying memory techniques. The possible damage from an Object injection or from an SQL injection or from is very high. Object injection could lead to remote code execution. An SQL injection may expose huge amounts of structured data, which is proven to be more valuable than raw data. Well formed structured data is easy to read, sort, search, and make sense of it. Via an SQL injection, an attacker could modify a database – insert, update, delete data, execute admin operations, recover file content, and even issue OS commands [25].

6.5 Next Most Dangerous CWEs

The next most dangerous groups of weaknesses in the MSSW class and base lists relate to file input and upload, authentication, randomization, cryptography, arithmetics and conversion, and input validation:

- randomization – class CWE-330 (Use of Insufficiently Random Values) with rank 5 is the class mostly directly assigned to CVEs.
- authentication – base CWE-798 (Use of Hard-coded Credentials) has rank 7; it is one of the contributors to the class CWE-287 (Improper Authentication) with the same rank 7 in the class list.
- file upload – base CWE-434 (Unrestricted Upload of File with Dangerous Type) has rank 8. It is the main contributors to class CWE-669 with rank 3.
- cryptography – base CWE-352 (Cross-Site Request Forgery) has rank 10, which relates to bugs in data verification. The class list also has class CWE-326 (Inadequate Encryption Strength) with rank 18, which is directly assigned to 35 CVEs with severity 7.24.
- arithmetics and conversion – base CWE-190 (Integer Overflow or Wraparound) and base CWE-191 (Integer Underflow) have ranks 13 and 20. They are the primary contributors to pillar CWE-682 (Incorrect Calculation) with rank 9. Others in this group on the top lists are bases CWE-131 (Incorrect Calculation of Buffer Size), CWE-190 (Integer Overflow or Wraparound), and CWE-191 (Integer Underflow – Wrap or Wraparound).
- input validation - base CWE-129 (Improper Validation of Array Index) has rank 16.

6.6 Mapping Dependencies

Both the MDSE and MSSW rankings heavily depend on how NVD assigns CWEs to particular CVEs. The CWE selection is restricted to view CWE-1003. Insufficient information about a CVE or an insufficiently specific CWE may lead to the use of the closest matching CWE class or pillar to describe the CVE. For example, it makes sense for class CWE-119 to be used for the memory corruption CVE-2019-7098, as there is not much information (no code and no details) – it could be any memory use error or a double free. However, there does exist enough information about the use after free CVE-2019-15554, but it is still mapped to class CWE-119 because there exists no appropriate base CWE. A close base CWE is CWE-416 (Use After Free), but it does not really reflect memory safe languages like Rust. It is also possible for a class CWE to be assigned to a CVE even when a specific base CWE is available. For example, the stack buffer overflow write CVE-2019-14363 is assigned class CWE-119, although there is plenty of information to map it more specifically to bases CWE-121 and CWE-120.

7 RELATED WORK

The constant need to improve information security has motivated a widespread interest in metrics (both qualitative [8] and quantitative [23]). As stated by Lord Kelvin, *you cannot improve if you cannot measure*. However, many members of the software security community doubt our ability to quantify security. Bellovin was among the first [2] to argue about the infeasibility of software security metrics. [4] discusses the limitations of the celebrated "Risk = Threat × Vulnerability × Consequence" model that is widely used. In [30] Verendel presents a critical survey of results and assumptions made in the community to quantify security. After reviewing over 100 articles, he concludes that the validity of most methods is still strikingly unclear. Many reasons explain this invalidity: lack of validation, lack of comparison against empirical data, and the fact that many assumptions in formal treatments are not empirically well-supported in operational security.

Although we agree, we posit that acceptable but possibly imperfect metrics must be developed in order to facilitate security decisions and to evaluate changes in security posture. To this end, there have been substantial efforts to produce security metrics; [30] surveys the literature of security metrics published between 1981 and 2008. More efforts can be found in [28], [26], and [20]. Security metrics that produce lists of the top security issues are also very prevalent [29], [11]. Specific to software security, there is the OWASP Top 10 [24] for web applications. Also, the CWE project has the Common Weaknesses Scoring System (CWSS) [14] and the Common Weakness Risk Analysis Framework (CWRAF) [16], which are used together to provide the most important weaknesses tailored to a particular organization.

There is also work to critique and improve the foundational data structures used by the MDSE and MSSW metrics. CWEs have been discussed in [31]. An entirely new approach to classifying software bugs (weaknesses) is proposed by [3] and is currently under development. The evolution of CWE is documented in [19] (e.g., the addition of classification trees and content for mobile applications and hardware). A critique of CVSS is available in [9]. In [12] a novel CWE data collection method is proposed along with simple atomic software security metrics. Our approach in contrast is an aggregate metric designed to be a direct replacement for the MDSE equation.

Along with much other work, our research should be considered as an important step in the process to improve CWE. We believe that our contribution is major as it points out a serious bias in the CWE MDSE equation that is preventing accurate measurements of the most significant software security weaknesses.

8 FUTURE WORK

This goal of this work is to identify and fix the unintended bias in the MDSE equation towards frequency. Thus we design the MSSW equation to, as evenly as possible, factor together frequency and severity. And this is rational as it models typical security risk matrices that equally combine probability and impact (e.g., [5]). However, it is possible that intentionally biasing towards either frequency or severity is more useful in this domain. Also, the CVSS severity equation is itself an aggregate of exploitability and impact. Future work should evaluate whether or not any intentional bias should be added between these 3 factors.

Also, future work should evaluate additional metrics that might be useful for determining the most significant CWEs. In particular, it would be useful to identify CWEs whose associated vulnerabilities are frequently used in actual and impactful breaches. We note that the CVSS temporal equations provide some of this, but these results are not commonly calculated and no public repository of this data exists. That said, some data does exist to support such mappings (e.g., [27]).

9 CONCLUSION

The field of security metrics is a difficult area of scientific research because there is often no ground truth, unlike disciplines such as physics and chemistry. This may lead one to focus on just taking simple low level measurements that are inherently defensible; that was the approach taken in [12]. However, creating aggregate metrics that compose multiple simple measurements is of practical importance for the field of security. In this work we did just that, aggregating frequency and severity (i.e., exploitability and impact) into a single metric. Our objective is not for the correlations to necessarily be equal, but that there exists a strong correlation for both factors which more evenly balances the inclusion of the top frequency and top severity CWEs. This seemingly simple task proved challenging because of the differing distributions of both simpler metrics. Indeed, the officially published CWE metric neglected this property and did not achieve its stated objective (almost exclusively choosing the most frequent CWEs). With our work, we claim to have addressed the limitations and to have produced the most accurate equation yet for measuring the most significant software security weakness.

REFERENCES

[1] David W Baker, Steven M Christey, William H Hill, and David E Mann. 1999. The Development of a Common Enumeration of Vulnerabilities and Exposures. In *Recent Advances in Intrusion Detection*, Vol. 7. Online proceeding, Purdue, IN, USA, 9.

[2] Steven M. Bellovin. 2006. On the Brittleness of Software and the Infeasibility of Security Metrics. *IEEE Security and Privacy* 4, 4 (July 2006), 96. https://doi.org/10.1109/MSP.2006.101

[3] Irena Bojanova, Paul E Black, Yaacov Yesha, and Yan Wu. 2016. The Bugs Framework (BF): A Structured approach to express bugs. In *2016 IEEE International Conference on Software Quality, Reliability and Security (QRS)*. IEEE, IEEE Press, Vienna, Austria, 175–182.

[4] Louis Anthony (Tony) Cox, Jr. 2008. Some Limitations of "Risk = Threat × Vulnerability × Consequence" for Risk Analysis of Terrorist Attacks. *Risk Analysis* 28, 6 (2008), 1749–1761. https://doi.org/10.1111/j.1539-6924.2008.01142.x

[5] Pamela A Engert and Zachary F Lansdowne. 1999. Risk matrix user's guide. *Bedford, MA: The MITRE Corporation* (1999).

[6] FIRST. 2019. Common Vulnerability Scoring System Special Interest Group. https://www.first.org/cvss Accessed: 2019-12-10.

[7] FIRST. 2019. Common Vulnerability Scoring System v3.1: Specification Document. https://www.first.org/cvss/v3.1/specification-document Accessed: 2020-2-5.

[8] Debra S. Herrmann. 2007. *Complete Guide to Security and Privacy Metrics: Measuring Regulatory Compliance, Operational Resilience, and ROI* (1st ed.). Auerbach Publications, USA.

[9] Allen D. Householder Art Manion Deana Shick Jonathan Spring, Eric Hatleback. 2018. Towards Improving CVSS. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=538368 Accessed: 2020-05-11.

[10] Robert A. Martin and Sean Barnum. 2008. Common Weakness Enumeration (CWE) Status Update. *Ada Lett.* XXVIII, 1 (April 2008), 88–91. https://doi.org/10.1145/1387830.1387835

[11] McAfee. 2020. McAfee Labs 2019 Threats Predictions Report. https://www.mcafee.com/blogs/other-blogs/mcafee-labs/mcafee-labs-2019-threats-predictions/ Accessed: 2020-02-01.

[12] Peter Mell and Assane Gueye. 2020. A Suite of Metrics for Calculating the Most Significant Security Relevant Software Flaw Types. In *2020 Conference on Computers, Software and Applications (COMPSAC)*. IEEE, IEEE Computer Society Press, Madrid, Spain.

[13] MITRE. 1999. Common Vulnerabilities and Exposures. https://cve.mitre.org Accessed: 2020-2-5.

[14] MITRE. 2018. Common Weakness Scoring System (CWSS). https://cwe.mitre.org/cwss/ Accessed: 2020-04-10.

[15] MITRE. 2019. Common Weakness Enumeration. https://cwe.mitre.org Accessed: 2019-12-10.

[16] MITRE. 2019. Common Weakness Risk Analysis Framework (CWRAF). https://cwe.mitre.org/cwraf/ Accessed: 2020-04-10.

[17] MITRE. 2020. 2019 CWE Top 25 Most Dangerous Software Errors. https://cwe.mitre.org/top25/archive/2019/2019_cwe_top25.html Accessed: 2020-02-01.

[18] MITRE. 2020. CWE Glossary. https://cwe.mitre.org/documents/glossary/ Accessed: 2020-05-11.

[19] MITRE. 2020. History of the Common Weakness Scoring System (CWSS). https://cwe.mitre.org/about/history.html Accessed: 2020-04-10.

[20] Patrick Morrison, David Moye, Rahul Pandita, and Laurie Williams. 2018. Mapping the field of software life cycle security metrics. *Information and Software Technology* 102 (2018), 146 – 159. https://doi.org/10.1016/j.infsof.2018.05.011

[21] NIST. 2020. BF Memory Model. https://samate.nist.gov/BF/Classes/MEMModel.html Accessed: 2020-05-11.

[22] NVD. 2020. National Vulnerability Database. https://nvd.nist.gov Accessed: 2020-01-10.

[23] Xinming Ou and Anoop Singhal. 2011. *Quantitative security risk assessment of enterprise networks*. Springer-Verlag, New York, NY, USA.

[24] OWASP. 2020. OWASP Top Ten. https://owasp.org/www-project-top-ten/ Accessed: 2020-04-10.

[25] OWASP. 2020. SQL Injection. https://owasp.org/www-community/attacks/SQL_Injection Accessed: 2020-05-11.

[26] Marcus Pendleton, Richard Garcia-Lebron, Jin-Hee Cho, and Shouhuai Xu. 2016. A Survey on Systems Security Metrics. *ACM Comput. Surv.* 49, 4, Article 62 (dec 2016), 35 pages. https://doi.org/10.1145/3005714

[27] Guy Podjarny. 2017. Which of the OWASP Top 10 Caused the World's Biggest Data Breaches? https://snyk.io/blog/owasp-top-10-breaches/ Accessed: 2020-09-22.

[28] T. W. Purboyo, B. Rahardjo, and Kuspriyanto. 2011. Security metrics: A brief survey. In *2011 2nd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering*. IEEE, Bandung, Indonesia, 79–82.

[29] Symantec. 2020. 2019 Internet Security Threat Report. https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf Accessed: 2020-02-01.

[30] Vilhelm Verendel. 2009. Quantified Security is a Weak Hypothesis: A Critical Survey of Results and Assumptions. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop* (Oxford, United Kingdom) *(NSPW '09)*. Association for Computing Machinery, New York, NY, USA, 37–50. https://doi.org/10.1145/

1719030.1719036

[31] Y. Wu, Irena Bojanova, and Y. Yesha. 2015. They know your weaknesses - Do you?: Reintroducing Common Weakness Enumeration. *CrossTalk* 28 (01 2015), 44–50.

SERIAL MICROFLUIDIC CYTOMETRY WITH INERTIAL AND HYDRODYNAMIC FLOW FOCUSING

Matthew DiSalvo[1,2], Paul N. Patrone[2], and Gregory A. Cooksey[2]
[1]Johns Hopkins University, USA and
[2]National Institute of Standards and Technology, USA

ABSTRACT

Microfluidics are increasingly used to develop flow cytometers with novel functionalities. Although various approaches exist to control particle positioning within microfluidics, the magnitude of and mechanisms leading to measurement uncertainties that arise from particle positioning within microfluidic channels remains an open question. Here, we report a novel combination of 3D-hydrodynamic and inertial focusing and demonstrate particle confinement, velocimetry, and quadruple replicate fluorescence measurements in a serial cytometer. Compared to inertial focusing alone, this combined focusing approach demonstrated lower particle velocity coefficient of variation (CV) (0.3 %), particle population fluorescence CV (9.4 %), and replicate measurement CVs (1.4 %).

KEYWORDS: Flow Cytometry, Microfluidics, Optofluidics, Reproducibility

INTRODUCTION

Flow cytometry is a primary tool for characterizing heterogeneous cell populations. Nonetheless, traditional cytometers are limited to a single measurement event per object and thus cannot assess the measurement uncertainty of each object. Recently, we reported an optofluidic serial cytometer that was capable of quadruple fluorescence measurements but was limited in throughput by particle velocities, which were below 1 mm s^{-1} [1]. The current work improves the consistency of the laser induced fluorescence by adding laser collimators and a combined inertial (IF) and 3D hydrodynamic focuser (3DHDF) to the design. By focusing to a single inertial node, serial measurements were recorded with a precision of 1.4 % at particle velocities above 0.3 m s^{-1}.

EXPERIMENTAL

(DISCLAIMER: Identification of commercial products does not imply recommendation or endorsement by NIST. The materials and equipment used may not necessarily be best for purpose.) Two-layer microfluidic devices were fabricated from poly(dimethylsiloxane) using soft lithography as reported [1], [2]. Combined IF&3DHDF utilized flow rates of (1, 4, 4, 10, and 40) µL min^{-1} for the sample stream, lower sheath, upper sheath, left sheath, and right sheath, respectively (See Figs. 1, 2A). The IF only control consisted of 59 µL min^{-1} sample flow. Four-channel digitized intensity spectra from photomultiplier tubes (Hamamatsu H11903-20) were recorded at a temporal resolution of 500 ns. Green fluorescent microspheres (15 µm diameter) were used for all experiments.

RESULTS AND DISCUSSION

The microfluidic serial cytometer performed four measurements of particle fluorescence: two replicate measurements per excitation, reproduced at two laser excitation regions spaced 1.2 cm apart (Fig. 1). When particles flowed at a particle-based Reynolds number $Re_p \approx 1.3$ within the 40 µm wide × 80 µm tall channels, they focused to two spatial nodes with maximum fluorescent intensities displaced approximately 9 µm from the centerline (Figs. 2B,C). This phenomenon is consistent with inertial focusing (IF) effects at similar aspect ratios and Re_p [3], and could lead to measurement variation due to varying the geometry between the laser, particle, and detector. To eliminate one inertial node and accelerate inertial equilibrium, particles were introduced at $Re_p \ll 1$ and then surrounded by a 3D water sheath biased 1:4 towards one side of the channel at the onset of $Re_p \approx 1.3$ flow (Fig. 2A). As characterized by microscopy image analysis, the focusing approach isolated microbeads to one node with a purity of 100 %, and variations in their positions were near the resolution limit of 0.23 µm per camera pixel (Fig. 2C).

Serial fluorescence and velocimetry measurements were obtained and compared between IF&3DHDF or Re_p-matched IF conditions to characterize the effect of the focusing approach. Measurements were recorded from 1536 green fluorescent microspheres over 3.9 minutes under IF&3DHDF and from 1475 particles over 2.4 minutes under IF. The four serial measurements for each particle were matched on the basis of anticipated time-of-flight, providing an analysis yield exceeding 97 % and 87 % for IF&3DHDF and IF conditions, respectively. Coefficient of variations (CVs) of integrated fluorescence intensity measurements generally decreased compared to the IF control (Fig. 3A). Particle velocity CVs decreased from 0.7 % to 0.3 % for IF and IF&3DHDF, respectively. After combining all 4 replicates, the population CV was 9.4 % with a precision (median of the replicate CV) of 1.4 % (Fig. 3B).

DiSalvo, Matthew; Patrone, Paul; Cooksey, Gregory A. "Serial microfluidic cytometry with inertial and hydrodynamic flow focusing."
Presented at 24th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS 2020). October 04, 2020 -
October 09, 2020.

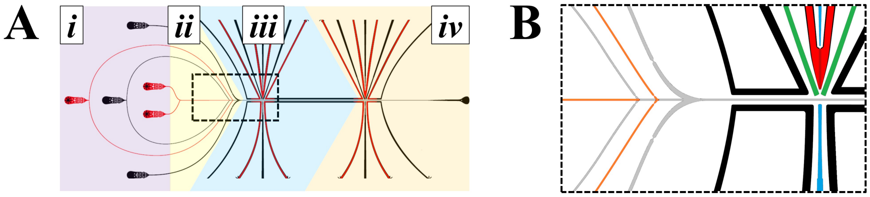


Figure 1: Microfluidic serial flow cytometer schematic. A) Device overview: i, inlets and particle traps; ii, flow focusing region; iii, measurement region 1; iv measurement region 2. B) Expanded view from the dashed area from A). Orange and gray: fluidic microchannels; black: light blocks; blue: waveguides; green: detection waveguides; red: light collimator.

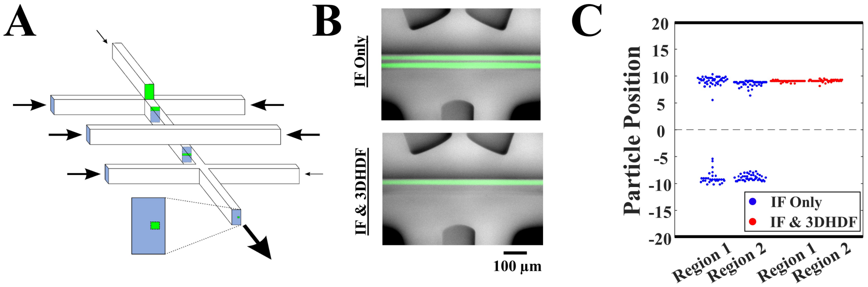


Figure 2: Inertial focusing (IF) and 3D hydrodynamic focusing (3DHDF) results. A) Principle of IF&3DHDF approach. B) Widefield composite microscopy images of single-particle streamlines. Green: particle fluorescence; Grayscale: bright-field. C) Distribution of particle positions. Dashed line: channel centerline; Solid lines: channel edges.

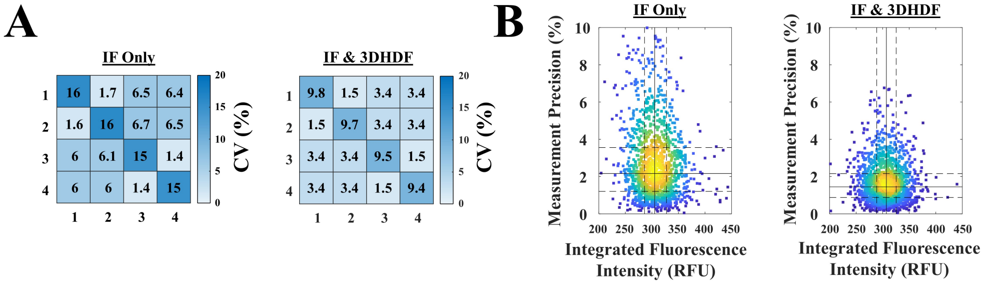


Figure 3: Integrated fluorescence area measurement precision. A) Heatmaps showing the pairwise comparison of CVs of integrated fluorescence intensities for the four replicated measurements. Diagonal: CVs of fluorescence intensities within one replicate. B) Scatter plots of measurements with associated measurement precision. Each point represents the mean fluorescence measurement from four replicates (x-axis) and the CV of those replicates (y-axis). Points are colored from blue to yellow in increasing density. Solid lines: median for each dimension; dashed lines: upper and lower quartiles.

CONCLUSION

Particle focusing methods and operating parameters can influence both the position and velocity of particles in flow cytometers, which impacts measurement repeatability. A serial cytometer provides a novel approach to directly observe measurement variations, enabling dissection of individual contributors to system uncertainty, thus elucidating paths towards maximizing measurement precision. In the future, the functionality of the serial cytometer can be extended by adding multicolor fluorescence and scattered light measurements. In addition, a systems control approach could be developed to adjust operating conditions on-the-fly to maximize precision over time.

REFERENCES

[1] G. A. Cooksey, P. N. Patrone, N. Podobedov, S. E. Meek, and J. A. Hsu, in *23rd Int. Conf. Miniaturized Syst. Chem. Life Sci*, 2019, October, 1508.

[2] G. A. Cooksey, P. N. Patrone, J. R. Hands, S. E. Meek, and A. J. Kearsley, *Anal. Chem.*, 91, 16, 10713–10722, Aug. 2019, doi: 10.1021/acs.analchem.9b02056.

[3] H. Amini, W. Lee, and D. Di Carlo, *Lab Chip*, 14, 15, 2739–2761, 2014, doi: 10.1039/c4lc00128a.

CONTACT: G.A. Cooksey; phone: +1-301-975-5529; gregory.cooksey@nist.gov

DiSalvo, Matthew; Patrone, Paul; Cooksey, Gregory A. "Serial microfluidic cytometry with inertial and hydrodynamic flow focusing."

MATCHING AND COMPARING OBJECTS IN A SERIAL CYTOMETER

Nikita Podobedov[1,2], Matthew DiSalvo[2,3], Jason Hsu[2,4], Paul Patrone[2], and Gregory A. Cooksey[2]
[1]Columbia University, [2]National Institute of Standards and Technology (NIST),
[3]Johns Hopkins University, [4]Montgomery Blair High School; USA

ABSTRACT

Flow cytometers are indispensable for clinical studies yet are limited by inherent uncertainties. We have developed an optofluidic device capable of multiple measurements along a microfluidic channel, whereby many of the uncertainty components can be systematically evaluated and reduced. This study addresses the challenge associated with identifying and tracking the order of objects throughout their time of transit. An algorithm was developed to characterize objects as they travel. We discuss methods of testing the efficacy of this algorithm, and other tools that allow us to gain more information than is provided by a conventional cytometer.

KEYWORDS: Flow Cytometry, Microfluidics, Optofluidics, Reproducibility

INTRODUCTION

Traditional flow cytometers can measure thousands of cells per second, but the intensity of the fluorescent biomarkers on cells is only measured once and used to calculate scalar values such as integrated area.[1] Thus, cytometers have limited ability to characterize biomarker distributions with high precision, which reduces their capacity to discriminate small changes within a population and to distinguish rare objects. Our group is developing an optofluidic cytometer that increases measurement resolution and precision by repeating measurements four times along the flow path: two axially symmetric waveguides collect fluorescence at each of two measurement regions separated along a microchannel. In order to aggregate the replicate measurements, an algorithm was created in MATLAB (MathWorks) to perform the piecewise temporal alignment across all four data channels. The algorithm evaluates the likelihood of object ordering based on time-of-flight and shape of the fluorescence signal, which includes high-temporal resolution of the intensity as the particle moves through the measurement region. We discuss challenges associated with objects passing one another, scenarios with multiple particles in the measurement region, such as doublets, and the unique opportunities that our high-resolution approach provides in extracting detailed information from objects, which would normally be impossible with other approaches. We also discuss metrics to assess the efficacy of matching and doublet separation.

EXPERIMENTAL

(DISCLAIMER: Identification of commercial products does not imply recommendation or endorsement by NIST. The materials and equipment used may not necessarily be best for purpose.) The manufacturing process for these devices has been detailed in a previous publication.[2] Briefly, two poly(dimethylsiloxane) (PDMS) layers with lithographed microchannels were aligned and bonded. The two measurement regions each included one waveguide which transmitted laser light and two symmetrically angled waveguides that collected emitted fluorescence from objects (15 µm diameter microspheres, Bangs Lab FSDG009) passing the laser (Fig. 1A). Fluorescence emission was measured on a photomultiplier tube (Hamamatsu H11903-20) and digitized (National Instruments PCIe-6374).

RESULTS AND DISCUSSION

Fluorescent microspheres passing through two measurement regions (each with a fluorescence collector angled upstream and downstream of the excitation) created signals as shown in Fig. 1. Calibration of our serial cytometer requires that standard objects be compared between regions to establish a scale factor for normalization. Importantly, this exercise requires unambiguous object matching from one region to another. Object matching was strongly dependent on a consistent particle velocity, which we maintained to within 0.2 % coefficient of variation (CV) using a combination of inertial and hydrodynamic focusing. Coincident events (doublets) impose additional challenges of identification and decoupling. Because our system records an object’s intensity both in time and from different perspectives at two different regions (Fig. 1B), a number of methods were tested in MATLAB, including idealized peak fitting, to facilitate individual object resolution and ordering (Fig. 1D, E).

To test algorithm efficiency, a number of *in silico* simulations were performed. Experimental data were lowpass filtered, normalized, temporally aligned, and averaged to form an idealized signal. We then modeled distributions of variables such as particle velocity, time between particles, white noise, and particle intensities from these data.

Podobedov, Nikita; DiSalvo, Matthew; Hsu, Jason; Patrone, Paul; Cooksey, Gregory A. ”Matching and comparing objects in a serial cytometer.” Presented at 24th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS 2020). October 04, 2020 - October 09, 2020.

Lastly, idealized signals were scaled, time-shifted, and had noise added to them to match the modeled distributions of those components, thereby creating simulated data equivalent to that of a mock bead population. Overall, when analyzing artificial signals, our algorithms correctly determined the identity of synthetic events across replicate measurement channels with 100 % tracking efficiency. In extracting integrated fluorescence area from the artificial signals, our algorithms demonstrated a precision of ± 0.04 % (± 1 standard deviation) compared to the synthetic ground truth. Using hydrodynamic focusing to position particles in the center of the channel, with event rates near what is typically achieved by commercial cytometers (1000 events per second), over 10 % of objects were predicted to pass each other between regions and ≈ 25 % could manifest as doublets (Fig. 1F). Adjusting hydrodynamic focusing to align particles to a single inertial focusing node reduced the velocity distribution to 0.2 % CV, leading to a 10-fold improvement in passing probability and 2-fold improvement in doublet likelihood.

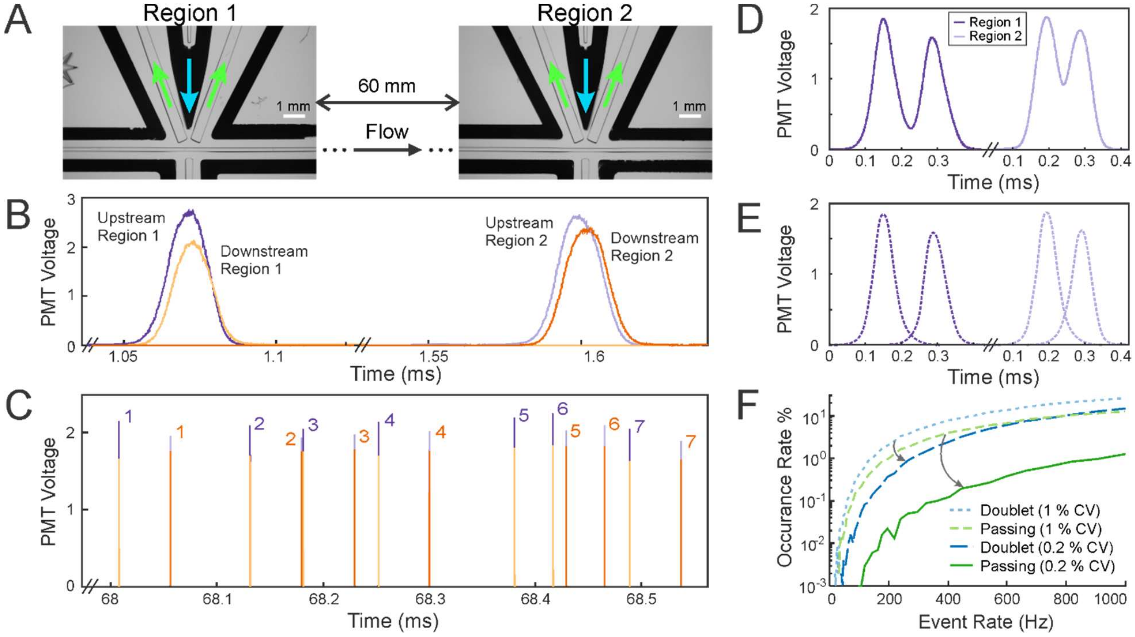


Figure 1: (A) Microscopy images of the 2 measurement regions in the microfluidic flow cytometer. Cyan: laser excitation; Green: fluorescence emission. (B) Representative fluorescence peaks as a bead travels through the 2 measurement regions. (C) Representative bead matching from time-of-flight and shape analysis. Numbers show tracking of each bead's data at region 1 (purple) and region 2 (orange). (D) Example of measured doublet and (E) simulated separation of each bead from the signal in both regions. (F) Simulation results comparing average rates of objects entering the device and the occurrence of passing and doublet events before (1 % velocity CV) and after (0.2 % velocity CV) improvement in flow focusing using combination of hydrodynamic and inertial focusing.

CONCLUSION

By providing multiple high-resolution measurements of objects in flow, our serial flow cytometer enables characterization of particles while promising improved discrimination of aberrant signal events. Importantly, we have demonstrated the need for an effective tracking and fitting algorithm for any microfluidic system that implements repeated measurement of objects along a flow path. Work is ongoing to extract additional measurement features from the upstream and downstream projections on the same object, including how these measurements can be used to extract the particle's size, shape, and fluorophore distribution.

REFERENCES

[1] H.M. Shapiro, "Practical Flow Cytometry, 4th edition," John Wiley & Sons Inc., New York, 2003.
[2] G.A. Cooksey, P.N. Patrone, et al., "Dynamic Measurement of Nanoflows: Realization of an Optofluidic Flow Meter to the Nanoliter-per-Minute Scale". *Anal. Chem.*, 91 (2019), pp. 10713-10722.

CONTACT

* G. Cooksey; phone: +1-301-975-5529; gregory.cooksey@nist.gov

A Preliminary Study on Uncertainty of NB-IoT Measurements in Reverberation Chambers

Anouk Hubrechsen[1,2], Vincent T. Neylon[2], Kate A. Remley[2], Robert D. Jones[2], Robert D. Horansky[2],
and Laurens A. Bronckers[1]

[1] Eindhoven University of Technology, Department of Electrical Engineering, Eindhoven, The Netherlands
[2] National Institute of Standards and Technology, Boulder, CO, USA

***Abstract*—New protocols related to internet-of-things applications may introduce previously unnoticed measurement effects due to the narrowband nature of these protocols. Such technologies also require less loading to meet the coherence bandwidth conditions, which may lead to higher variations accross the channel. This can cause a need to take additional components into account in the assessment of uncertainty. In this work, we present a preliminary study on uncertainties of NB-IoT measurements in reverberation chambers. We show a need to account for both the number of mode-stirring samples and the lack of spatial uniformity in the uncertainty analysis, where the latter generally dominates for wireless testing. We provide preliminary results for the uncertainty including both effects. We introduce a hypothesis for the effects of loading on the uncertainty, introducing that there may be an optimal loading point to minimize uncertainty, where we describe that this decision may not depend only on coherence bandwidth, but also on the number of significant modes.**

***Index Terms*—CAT-M1, Cellular Telecommunications, Chamber transfer function, Internet of Things, Mode distribution, NB-IoT, Reverberation chamber, Uncertainty, Wireless System**

I. Introduction

The use of internet-of-things (IoT) or machine-to-machine (M2M) applications is gaining popularity to meet demands such as ubiquitous coverage, increased reconfigurability, and mobility, that are required for 5G and beyond [1], [2]. These devices will largely work in the lower 5G, or sub-6 GHz, bands and will be using protocols such as the narrowband IoT (NB-IoT) and CAT-M1 (or LTE-M) protocols [1], [2].

The performance of these cellular devices is often studied over-the-air (OTA) by metrics such as Total Isotropic Sensitivity (TIS) and Total Radiated Power (TRP) [3]. These can be carried out either in an anechoic chamber (AC) or a reverberation chamber (RC). An RC is a large metal cavity, with one or more mode-stirring mechanisms to produce, on average, a uniform distribution of the fields, and can often produce faster, lower-cost, or more flexibly configurable measurements than an AC [4]. This makes an RC an excellent candidate for testing IoT devices.

RCs have been researched extensively and were shown to be suitable for TIS measurements on earlier-generation protocols, such as WCDMA (5 MHz channel bandwidth). However, for NB-IoT we expect additional challenges due to the narrowband nature of this protocol (180 kHz channel bandwidth). Traditionally, to provide accurate results, a wideband RC reference measurement is averaged over frequency in post processing to match the bandwidth of the modulated signal. Such frequency averaging has the added benefit of smoothing the frequency response. When averaging the frequency response over a narrow bandwidth, peaks and nulls in the RC's frequency response for the mode-stirring samples may increase uncertainty, as we will show.

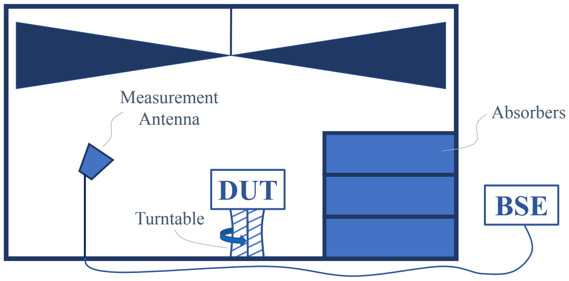


Fig. 1. Illustration of the RC setup for TIS, including a turntable for position stirring, which is needed in loaded chamber measurements.

Earlier research has extensively studied uncertainty effects in loaded RCs for wireless-device testing [5]–[7]. For the first time, we provide a preliminary study where we show a need to incorporate both the uncertainty due to the number of mode-stirring samples *within* a data set and the uncertainty due to a lack of spatial uniformity, captured *between* data sets. We introduce a hypothesis for the effects of loading on the uncertainties. In Section II we explain the process for TIS measurements in RCs. In Section III, we provide the methodology and experimental results, where we incorporated the *within* uncertainty. In Section IV, we introduce a hypothesis support the results. The work is concluded in Section V.

II. TIS Measurement

TIS is a measure of the minimum received power that a device can accept without incurring an unacceptable throughput for a certain protocol. An illustration of a typical RC setup for a TIS measurement is shown in Fig. 1. A wireless link is set up between a base-station emulator (BSE) and a device under test (DUT), where the BSE transmits a signal at decreasing power levels at the downlink frequency, and measures the DUT's

Hubrechsen, Anouk; Neylon, Vincent; Remley, Kate; Jones, Rob; Horansky, Rob; Bronckers, Laurens. "A Preliminary Study on Uncertainty of NB-IoT Measurements in Reverberation Chambers." Presented at European Microwave Week (EuMW 2020). January 10, 2021 - January 15, 2021.

reported throughput at the uplink frequency. TIS is defined as the minimum power incident on a DUT where the throughput drops below a certain percentage. For the NB-IoT protocol, a throughput of 95 % is used. We measure the BSE power for a high value of starting power and as long as the throughput is 95 % or higher [8], we step the power down until the throughput drops below 95 % to obtain a minimum power for each mode-stirring sample. This process is repeated for every sample in the mode-stirring sequence, and then averaged over all mode-stirring samples to obtain TIS [8].

Usually, we need to load the chamber by adding RF absorbers to flatten the RC's frequency response and to keep the communication link between the BSE and the DUT while measuring TIS. This is due to the fact that, in an unloaded chamber, the frequency selectivity is usually too high for the DUT's equalizers. This increases frequency correlation and reduces spatial uniformity, which may increase uncertainty if not compensated for using position stirring with, for example, a turntable as shown in Fig. 1 [5]. The amount of loading necessary can be determined using the coherence bandwidth (CBW), defined by the average bandwidth over which the frequency samples have a minimum specified level of correlation [9]. In general, the CBW needs to be wider than the channel bandwidth to maintain the link [8].

In the CTIA Test Plan for Large-Form-Factor IoT Devices [8], TIS is calculated by

$$P_{\text{TIS}} = G_{\text{ref}}\eta_{\text{meas}}^{\text{tot}}G_{\text{cable}}(\langle\frac{1}{P_{\text{BSE}}}\rangle_M)^{-1}, \qquad (1)$$

where P_{TIS} is the total isotropic sensitivity in W and $\eta_{\text{meas}}^{\text{tot}}$ the total efficiency of the measurement antenna. G_{cable} is the cable loss between the measurement antenna and the BSE, $P_{\text{BSE}(m)}$ is the minimum received power measured by the BSE at the threshold throughput in W for mode-stirring sample m, $\langle\cdot\rangle_M$ is an ensemble average over the total number of mode-stirring samples M and G_{ref} is the chamber transfer function given by

$$G_{\text{ref}} = \frac{\langle\langle|S_{21}|^2\rangle_M\rangle_F}{\eta_{\text{meas}}^{\text{tot}}\eta_{\text{ref}}^{\text{tot}}}, \qquad (2)$$

where $\eta_{\text{ref}}^{\text{tot}}$ is the total efficiency of the reference antenna and $\langle\cdot\rangle_F$ is an ensemble average over F frequencies across the channel bandwidth. According to the standard [8], G_{ref} is used in the assessment of uncertainty, through the standard deviation of the measurement over several independent realizations of the mode-stirring sequence, defined as the *between* uncertainty. G_{ref} needs to be frequency averaged over the same bandwidth as the DUT channel being measured. However, since the channel bandwidth for NB-IoT is much narrower than previously used protocols, this will result in less frequency averaging so we may expect uncertainty to increase, as we show next.

III. Experimental Results

In this section, we first present our measurement setup and the methods we used for estimating uncertainty. Then, we show preliminary experimental results on uncertainty.

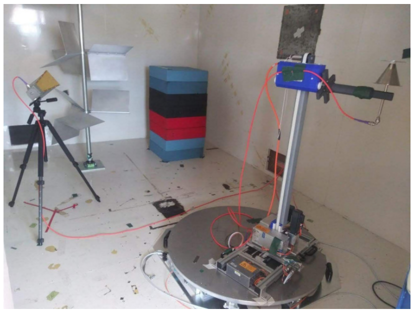


Fig. 2. RC setup to measure G_{ref} for eight absorbers. The chamber contains one vertical paddle for mode stirring, a turntable with height translation for position stirring, and automated polarization stirring.

TABLE I
Mode-stirring sequence for each independent realization (IR)

IR	Height	Pol	Paddle angles		Turntable Angles
			Angles	Offset	
1-3	0.3 m	0°	8	0°, 15°, 30°	15
4-6	0.3 m	90°	8	0°, 15°, 30°	15
7-9	1.3 m	0°	8	0°, 15°, 30°	15
10-12	1.3 m	90°	8	0°, 15°, 30°	15

A. Experiment Setup

Measurements were carried out in a 4.6 m x 3.1 m x 2.8 m RC at the National Institute of Standards and Technology (NIST), as shown in Fig. 2, which has one paddle as a mode-stirring mechanism, a turntable and height translation for position stirring, and automated polarization stirring. From all mode-stirring samples, we acquired 12 independent realizations ($N_B = 12$), each containing 120 mode-stirring samples ($N_W = 120$) *within* the stirring sequence obtained from 8 paddle and 15 turntable angles with 45°and 24°angle spacing, respectively, as shown in Table I. We state that they are independent realizations based on an application of cross-correlation techniques that verified independence to within a specified threshold of 0.3 [8]. A vector network analyzer (VNA) was used, with an IF BW setting of 1 kHz, a source power of -8 dBm and a 1 kHz frequency spacing over 3 different frequency bands of 10 MHz, centered at 1993 MHz, 1995 MHz and 1999 MHz in the Cellular NB-IoT Band 2. In post processing, we chose multiple frequency averaging bandwidths to observe the effect of a narrow averaging bandwidth across frequency. We also chose two absorber cases, one with two RF absorbers (CBW = 1.5 MHz) and one with eight (CBW = 3.3 MHz), where the CBW was calculated with a threshold of 0.5. We used two low-loss broadband antennas

TABLE II
Percentage of *between* significance in the frequency band

	180 kHz	1.2 MHz	2.0 MHz
2 Abs. (CBW ≈ 1.5 MHz)	8 %	15 %	28 %
8 Abs. (CBW ≈ 3.3 MHz)	31 %	39 %	53 %

Hubrechsen, Anouk; Neylon, Vincent; Remley, Kate; Jones, Rob; Horansky, Rob; Bronckers, Laurens. "A Preliminary Study on Uncertainty of NB-IoT Measurements in Reverberation Chambers." Presented at European Microwave Week (EuMW 2020). January 10, 2021 - January 15, 2021.

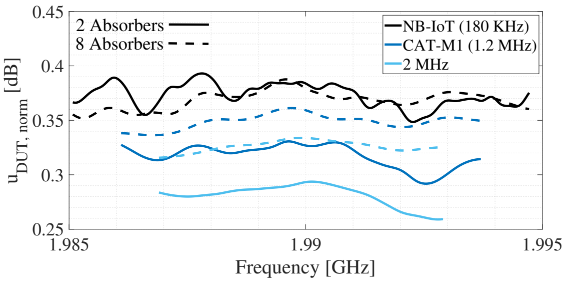


Fig. 3. Uncertainty of the DUT for two (CBW ≈ 1.5 MHz) and eight absorbers (CBW ≈ 3.3 MHz), calculated using the formulation where *within* uncertainties are dominant, for three different averaging bandwidths.

for the measurement, where the calibration reference plane was brought up to the connectors of the antennas using an N-type electronic calibration module. The measurement setup with eight absorbers is shown in Fig. 2.

B. Significance Test for Uncertainty

We perform a significance test as described in detail in [6] to determine if only the *between* uncertainty is significant, as is dictated by the current standard [8] or if the *within* uncertainty, defined by the variation due to the number of samples *within* an independent realization, should be taken into account as well, as different formulations are used depending on the outcome. As shown in Table II for the higher end of Band 2, the percentage of the frequency band where *between* uncertainties are significant increases for higher-loading cases, due to a reduced spatial uniformity. However, for both NB-IoT and CAT-M1 the *between* uncertainties are not significant for the majority of the frequency band, so the *within* component should be taken into account as well in the uncertainty (similar results were observed for the other two bands). This shows a need to reassess current uncertainty methods to be applicable to NB-IoT, which will be discussed more extensively in future publications. The formulation used in this work (taking *within* uncertainty into account as well) to determine the uncertainty of the DUT is given by [6]

$$u^2_{\mathrm{DUT}} = \frac{1}{N_W(N_B N_W - 1)} \sum_{i=1}^{N_W} \sum_{j=1}^{N_B} (G_R(w_i, b_j) - \hat{G}_{\mathrm{Ref}})^2. \tag{3}$$

Next, we show results using this formulation. Note that a similar effect, where the dominant source of uncertainty changes with the specific chamber configuration, was also observed in [5], [6], but with less focus on *within* uncertainties.

C. Experiment Results

Fig. 3 shows the normalized u_{DUT} for loading cases with two and eight absorbers, averaged over various channel bandwidths. For the sake of brevity, we only show the higher-end of Band 2, but similar results were observed for the mid- and lower-end. In the current CTIA Large-Form-Factor IoT Device Test Plan [8], the user selects the highest value of u_{DUT}, computed over all frequencies within the band of interest, to find uncertainty, since, as shown in Fig. 3, uncertainty estimates can change over frequency. There are two main findings in these results. First, the maximum uncertainty is similar for both absorber cases, while, generally, uncertainty increases for increased loading [9]. Second, an increased averaging bandwidth reduces the uncertainty more in the two-absorber case, as compared to the eight-absorber case. Both effects are generally not observed in RC measurements with a wider averaging bandwidth, although for wider bandwidths, the *between* uncertainty often dominates. Next, we introduce a hypothesis for these effects. Note that this is a preliminary study so other explanations may be valid as well. Measurements over multiple bands and absorber cases in different chambers should be performed, and compared to a TIS measurement, to form definite conclusions.

IV. Mode Distribution and Uncertainty

In this section, we introduce a possible explanation for the results, linking the amount of loading to changes in *within* and *between* uncertainties.

In an unloaded chamber, each mode is expected to be very narrowband compared to the channel bandwidth, as shown for an ideal chamber in [10]. When we load the chamber, these individual modes become wider, and, therefore, the frequency response flattens, increasing the CBW. This is illustrated in Fig. 4 (note that we normalized the maximums to illustrate a concept, it is not an illustration of the actual modes in our chamber). Considering that the NB-IoT protocol has a maximum channel bandwidth of 180 kHz, and that the CBW of an unloaded chamber is often wider than 180 kHz [9], it should be possible to perform the measurement with no loading. In wider-band measurements, this generally reduces *between* uncertainty, due to a higher spatial uniformity. However, earlier work on NB-IoT TIS measurements did use chamber loading [11], but no reason was stated. Besides that, the results show a similar maximum *within* uncertainty for

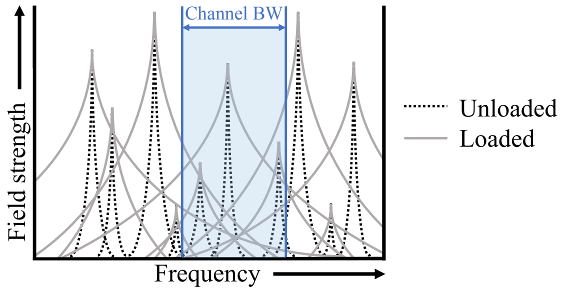


Fig. 4. Illustration of the mode distribution in a loaded and an unloaded RC. The modes spread more across frequency in the channel bandwidth for higher loading cases. Even though these are all correlated within the CBW, sharp peaks and nulls are flattened by the addition of more significant modes.

both loading cases. Therefore, a need arises to assess the effects of chamber loading on the *within* uncertainty together with the question if some chamber loading may be preferable, even in cases where the CBW is already wider than the channel bandwidth.

We expect the following effects to influence *within* uncertainty in cases when the channel bandwidth is narrow:

- A higher averaging bandwidth reduces the amount of peaks and nulls in the frequency response, reducing *within* uncertainty. Therefore, a user may require more mode-stirring samples for narrowband measurements. This effect can be observed in both CBW cases in Fig. 3.
- Fewer significant modes (modes with a significant contribution to the measured S-parameter at a given frequency [12]) may be included in each channel for narrowband measurements. This may result in higher peaks and nulls in the frequency response and an increased *within* uncertainty (this is likely related to the previous point).
- If chamber loading is increased, the frequency response flattens, and the amount of significant modes may increase (contributions from modes that were mostly outside of the channel bandwidth in an unloaded case may occur in-band for a loaded case) [12], [13], as illustrated in Fig. 4. Due to the presence of more significant modes within the band, the *within* uncertainty may decrease, even though all the frequency samples in the channel bandwidth are more correlated due to increased loading. Note that this example only holds for one mode-stirring sample, but consequently could affect the average of all measurements as well. This is a possible explanation on why increased frequency averaging had less of a reduction in uncertainty in the 8-absorber case, since *within* differences may have been reduced already by loading, reducing the effect of additional frequency averaging.

Combining these three statements on the effect of loading on the uncertainty as defined in (3), one could state that there may be an optimal-loading point for narrowband TIS measurements, as illustrated in Fig. 5. This point would be a trade-off, where the frequency response is flattened sufficiently such that peaks and nulls are averaged out, but the chamber is not loaded to a point where the spatial uniformity decreases in such way that *between* differences become significant, given a significant amount of position stirring. Note that a similar effect was observed in [13] for small added loss, where loss resulted in a benefit to convergence to an overmoded (sufficient significant modes) condition. The proposed approach could be more extensive than the current approach where one defines the amount of loading only by use of the CBW, but as stated earlier, a more extensive study should be performed to form definitive conclusions.

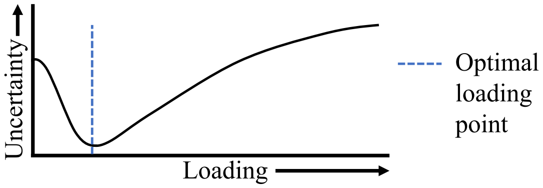


Fig. 5. Illustration of the expected effect on uncertainty by an increase in chamber loading. Uncertainty first decreases because there are more significant modes in the channel bandwidth, but increases for higher loading due to high correlations.

V. Conclusion

In this paper, we have shown that current methods in the assessment of uncertainty for TIS measurements of IoT devices may need to be reassessed due to the narrowband nature of the NB-IoT protocol. We introduced a hypothesis that links the effects of loading on the mode-distribution to a reduced uncertainty, taking both variations *between* and *within* the data sets into account, leading to a possible optimal-loading point. However, more extensive measurements should be performed and compared to the uncertainty of a full TIS measurement to form definite conclusions. Future publications will contain a more extensive assessment of the uncertainty in (3), and will provide more elaborate results for u^2_{DUT} and TIS.

References

[1] GSMA, “Mobile IoT in the 5G future: NB-IoT and LTE-M in the context of 5G,” Apr 2018.

[2] 3GPP TS 36.101, “Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception (release 16),” Dec 2019.

[3] R. D. Horansky, T. B. Meurs, M. V. North, C. Wang, M. G. Becker, and K. A. Remley, “Statistical considerations for total isotropic sensitivity of wireless devices measured in reverberation chambers,” in *2018 International Symposium on Electromagnetic Compatibility (EMC EUROPE)*, Aug 2018, pp. 398–403.

[4] R. D. Horansky and K. A. Remley, “Flexibility in over-the-air testing of receiver sensitivity with reverberation chambers,” *IET Microwaves, Antennas Propagation*, vol. 13, no. 15, pp. 2590–2597, 2019.

[5] M. G. Becker, M. Frey, S. Streett, K. A. Remley, R. D. Horansky, and D. Senic, “Correlation-based uncertainty in loaded reverberation chambers,” *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 10, pp. 5453–5463, Oct 2018.

[6] K. A. Remley, C. J. Wang, D. F. Williams, J. J. aan den Toorn, and C. L. Holloway, “A significance test for reverberation-chamber measurement uncertainty in total radiated power of wireless devices,” *IEEE Transactions on Electromagnetic Compatibility*, vol. 58, no. 1, pp. 207–219, Feb 2016.

[7] X. Chen, P. Kildal, and S. Lai, “Estimation of average rician k-factor and average mode bandwidth in loaded reverberation chamber,” *IEEE Antennas and Wireless Propagation Letters*, vol. 10, pp. 1437–1440, 2011.

[8] CTIA Certification, “Test plan for wireless large-form-factor device over-the-air performance, version 1.2.1,” Feb 2019.

[9] K. A. Remley, J. Dortmans, C. Weldon, R. D. Horansky, T. B. Meurs, C. Wang, D. F. Williams, C. L. Holloway, and P. F. Wilson, “Configuring and verifying reverberation chambers for testing cellular wireless devices,” *IEEE Transactions on Electromagnetic Compatibility*, vol. 58, no. 3, pp. 661–672, June 2016.

[10] P. Besnier and B. Démoulin, *Electromagnetic Reverberation Chambers*, 2011, vol. 1.

[11] J. Luo, E. Mendivil, and M. Christopher, “Over-the-air performance evaluation of NB-IoT in reverberation chamber and anechoic chamber,” in *2018 AMTA Proceedings*, Nov 2018, pp. 1–3.

[12] F. Monsef and A. Cozza, “Average number of significant modes excited in a mode-stirred reverberation chamber,” *IEEE Transactions on Electromagnetic Compatibility*, vol. 56, no. 2, pp. 259–265, 2014.

[13] A. Cozza, “The role of losses in the definition of the overmoded condition for reverberation chambers and their statistics,” *IEEE Transactions on Electromagnetic Compatibility*, vol. 53, no. 2, pp. 296–307, 2011.

**Proceedings of the ASME 2020
International Mechanical Engineering Congress and Exposition
IMECE2020
November 16-19, 2020, Portland, OR, USA**

IMECE2020-24312

INFORMATION MODEL FOR A-UGV PERFORMANCE MEASUREMENT STANDARD

Soocheol Yoon[1], Roger Bostelman
National Institute of Standards and Technology
Gaithersburg, MD

ABSTRACT

Automatic through autonomous - Unmanned Ground Vehicles (A-UGVs) have the potential to be applied over a wide range of manufacturing systems under industry 4.0 paradigm. In order to use A-UGVs efficiently in a manufacturing system, it is necessary to select the A-UGV suitable for each factory or workplace. A framework for evaluating the A-UGV performance under a specific manufacturing environment is needed. ASTM International Committee F45 has been developing standards[2] providing a basis for A-UGV manufacturers and users to compare tasks to the A-UGV capabilities. This paper proposes information models for A-UGV performance measurement along the standards development. The standard needs are analyzed to show how the standard and information model can be used for the introduction of A-UGVs into factories. The information model in this paper provides a structured way to describe the factory elements affecting the A-UGV performance, and the measured A-UGV performances against the factory elements. To validate the proposed information model, an A-UGV performance testbed was built and the information model instance is developed to describe the testbed elements. An A-UGV is tested against the testbed elements and the measured performance is described by the other instance. This paper contributes to mutual understanding, between A-UGV makers and users, to deliver A-UGV performance information efficiently and to provide basis for A-UGVs to be tested under the same conditions.

Keywords: autonomous mobile robots, performance standard, navigation, obstacle avoidance, information model

1. BACKGROUND

With the development of information technology (IT), driverless vehicles are being developed and applied in many ways. In the manufacturing domain, the development and application of various types of unmanned vehicles, including the well-known Automatic Guided Vehicle (AGV), have been continuously conducted [1,2]. In the meantime, the emergence of Automatic through Autonomous – Unmanned Ground Vehicles (A-UGVs) is the starting point for the manufacturing system to more dynamically respond to changes. A-UGV is defined as "automatic, automated or autonomous vehicle that operates while in contact with the ground without a human operator" [3]. In terms of logistics support, A-UGVs enable active logistics flow control by real-time scheduling and execution. Beyond logistics support, an A-UGV combined with a robot arm, called a mobile manipulator, can autonomously perform manufacturing processes like assembly. A-UGVs have the potential to implement digital manufacturing, smart factories, and industry 4.0 in various ways [4]. Autonomy is one of the keys to implement Industry 4.0, and A-UGVs can serve in advanced logistic and manufacturing applications. For example, the modular factory [5] aims at a plug-and-produce manufacturing system, which makes it easy to add or change the production line modules as needed. A-UGVs can be used to transport the modules making the manufacturing system more flexible and autonomous. Delivering manufacturing components is one of the areas that have been actively researched [6].

Every A-UGV application assumes that the A-UGV can move freely in the factory. In terms of the A-UGV movement, laboratories, aiming to develop and test new technologies, and factories, aiming to manufacture products, have completely different A-UGV operational focus. In the laboratory, A-UGVs are developed and tested where the experiments may not test the required vehicle performance needed in the factory—for example, low or bright lighting, floor surface defects, stationary or moving obstacles, etc. On the other hand, factories are focused

[1] Contact author: soocheol.yoon@nist.gov

[2] Commercial products are identified in this paper to foster understanding. This does not imply recommendation or endorsement by NIST, nor that the products identified are necessarily the best available for the purpose.

Yoon, Soocheol; Bostelman, Roger V. "Information Model For A-Ugv Performance Measurement Standard." Presented at International Mechanical Engineering Congress and Exposition (IMECE 2020). November 16, 2020 - November 19, 2020.

on the manufacturing process and the products they make, and A-UGVs are one of the tools they use to be more efficient and effective.

In order to introduce A-UGVs into a factory, it is necessary to understand the elements that consist of the factory and the expected A-UGV performance. However, the very different situations for the A-UGV maker and the user make it difficult to grasp. It is impossible for the A-UGV maker to test in advance in the laboratory, and for the A-UGV user (i.e., factory) to know, what affects the A-UGV performance in the actual manufacturing environment. At a minimum, it is necessary for makers and users to share A-UGV performance for conditions (environment and obstacles) that are commonly present in the factory.

Previous studies defined common environmental conditions and obstacles that should be identified for A-UGV use in the factory [7]. The effects of factory commonalities, including ramps, lighting, forklifts, and pedestrian walkways, and the required function and performance of A-UGV against the commonalities were analyzed. Each factor can have a significant effect on A-UGVs possibly causing accidents and needs to be managed.

Reflecting the conducted studies, ASTM International (formerly: American Society for Testing and Materials) F45 committee is developing driverless automatic guided industrial vehicles standards for environment, navigation, and obstacle detection[3]. ASTM F45 standards include A-UGV performance measurement and documentation methods for exact replication of vehicle tests. Through this, the user can define the A-UGV operating environment prior to implementation, and better express the expected A-UGV performance upon integrating the A-UGV into the factory. In addition, A-UGV manufacturers can build a mock test environment based on factory environment information, test A-UGVs according to measurement methods, and express A-UGV performance.

There has been research to define performance metrics, to develop measurement technologies and test methods. Bostelman et al [8] summarizes the existing studies for measuring the performance of A-UGVs and mobile manipulators. According to the summary, the metrics have been defined for the measurement of time/task duration, distance traveled, repeatability, accuracy, effectiveness, efficiency, autonomy, etc. The summary includes the measurement technologies like wireless indoor position measurement, perception sensing, motion tracking camera, laser tracking, and indoor global position system. It also includes measurement methods like benchmark, simulator, competition, and existing and ongoing international standards.

There have been many studies to enhance the performance, especially of object detection and navigation. Young et al [9] suggest framework to predict performances of A-UGV according to outdoor terrain. Wu et al [10] applied neural networks to single colored camera object detection.

Although there are research needs for describing the factors, and the A-UGV performance against them, in the factory, research to identify and define them have not been conducted yet. Among existing studies, it is difficult to determine how various obstacles or environments may affect the A-UGV's ability to navigate, to recognize obstacles, and to avoid obstacles. This means that the manufacturing environment combined with the user's requirements for the implementation of A-UGVs in factories have not been fully analyzed.

Information models are a good way to describe factory conditions and A-UGV performance information quickly and accurately. Unified Modeling Language (UML) [11] defines various diagrams to describe system components and behaviors. Among them, a class diagram well represents the components, properties, and relationships of the system, and is good for factory environment and vehicle performance information modeling. In addition, an object diagram, which is an instance of the class diagram, can be used to describe the measured A-UGV performance and the actual factory environment.

Accordingly, this paper aims to develop factory environment and vehicle information models for the introduction of A-UGVs into factories. The class diagram and object diagram of UML are used as modeling methods. The scope of this paper considers the environmental and obstacle factors including the elements discovered by previous research [7] and defined in the ASTM F45 standards [3]. In addition, newly discovered elements are added as well.

Before developing an information model, it is important to analyze why the standards and information models are needed based on the difference between the user and the maker. To this end, the requirements analysis comes first prior to information model development. Then, a factory information model for the introduction of A-UGVs, and A-UGV performance information models for evaluation, are developed. As a case study, an instance for a factory-like testbed and an A-UGV is developed.

2. NEEDS FOR A-UGV PERFORMANCE STANDARDS

Manufacturing standards generally serve as a guide for designers, engineers, builders, operators, and decision makers [12]. They also play roles in facilitating communication between stakeholders across manufacturing lifecycles, manufacturing supply chains, and from the factory floor to enterprise domains. In this paper, the focus is on the mutual understanding between A-UGV makers and users. This understanding will allow determination of whether the A-UGV can be operated or not and if so, performance estimation of A-UGVs within their operating environment. As a result, this will help efficient use and market expansion of A-UGVs. This section will first analyze the different views between A-UGV makers and users. Then, the section will suggest how standards can help enable mutual understanding between makers and users, and later, summarize what items should be defined as standards.

The analysis in this section shows one of the ways to use standards. The purpose of standards varies and is not limited to this.

2.1 Different views between makers and users

A-UGVs have various technical specifications. In catalogs, it is easy to find specs such as maximum speed, maximum load, runtime, charging method and time, navigation performance, etc.

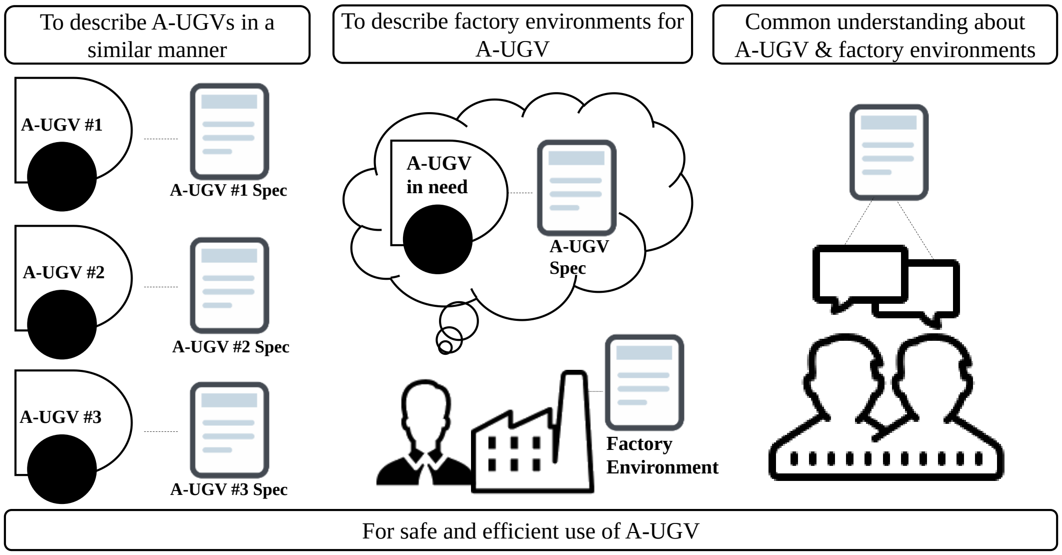

To describe A-UGVs in a similar manner
To describe factory environments for A-UGV
Common understanding about A-UGV & factory environments
A-UGV #1
A-UGV #1 Spec
A-UGV #2
A-UGV #2 Spec
A-UGV #3
A-UGV #3 Spec
A-UGV in need
A-UGV Spec
Factory Environment
For safe and efficient use of A-UGV

FIGURE 1: STANDARDS NEEDS SUPPORTING A-UGV USERS AND MAKERS

In addition, physical components specs such as drive wheel(s), motor, amplification, and battery can be numerically expressed. In terms of software, it shows what can be controlled including driving mode, driving function, map making, navigation, and monitoring. In terms of communication, it shows how the user can control the A-UGV including remote control, monitoring, fleet mode control, etc.

Meanwhile, there is less information about A-UGV performance under specific conditions. It is impossible for an A-UGV maker to test all manufacturing environments in the world, so they assume the general manufacturing environment and test the performance of the A-UGV in their facilities. However, this general environment may not equate well to factories in the world. Therefore, A-UGV makers need a method to express A-UGV performance under specific conditions.

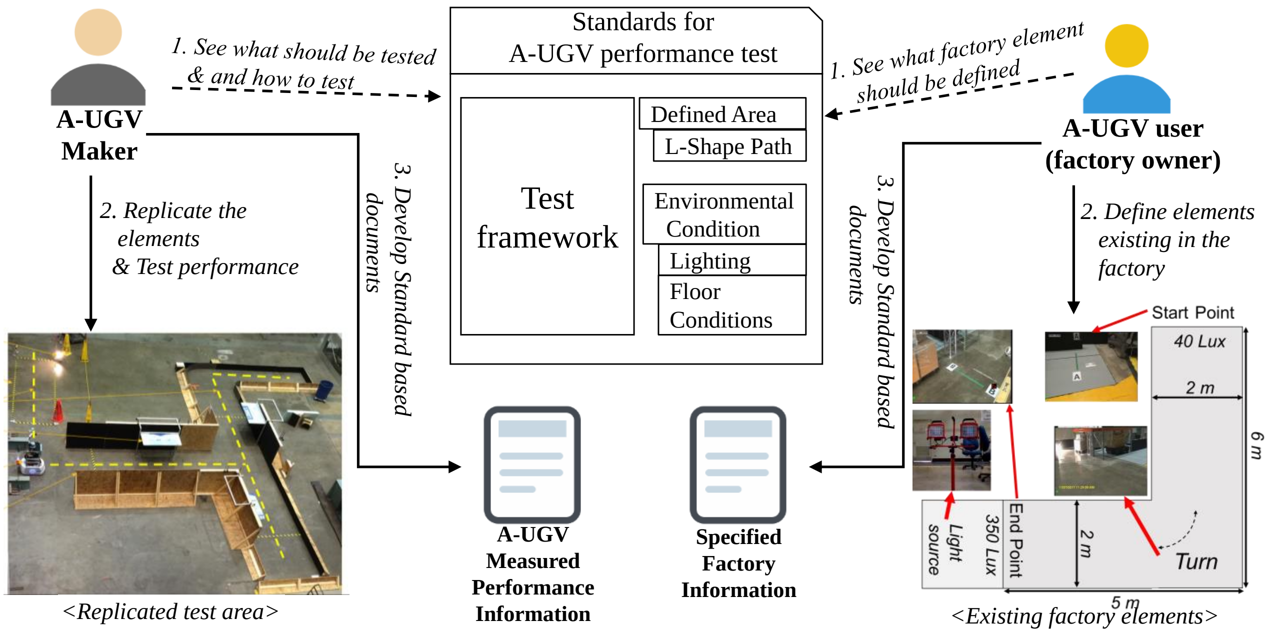
From the user side, the first thing to consider is whether A-UGVs can perform the task in their manufacturing environment, rather than just purchasing A-UGV's based on their mechanical specifications. Manufacturing environments include basics such as temperature, humidity, brightness or lighting conditions, energy supply, and floor conditions. Open or closed space, workspace size, obstacles, pedestrians, floor level transitions (e.g., ramps), doors, gates, or steps (e.g., thresholds) should be considered. Safety, health, and quality should be protected and maintained.

The information provided by a maker alone is not enough to determine whether A-UGVs can operate in the user's manufacturing environment. Factory workers are not typically A-UGV experts although they know their process. The A-UGV installer must use their expertise to integrate the vehicle into the facility. However, once the installer leaves, many issues can arise that affect the A-UGV performance where user-knowledge is gained only during A-UGV use, instead of cost- and time-saving A-UGV performance knowledge gained prior to the installation. A-UGV users therefore need a means to express how the environment affects the A-UGV performance.

Figure 1 depicts overall needs for standards consisting of one base objective and three detailed objectives to resolve the issues. The base objective upon introduction of A-UGVs is that they are used safely and efficiently. To do this, first, standards need to describe the A-UGVs performance using a standardized method so that all A-UGVs are similarly described. Example A-UGV performance standards are ASTM F3244 navigation [13] and WK57000 docking. The A-UGV performance can be described for the A-UGVs being made or as requirements from users. Next, standards need to describe the factory environment using a standardized method. Here, the factory environment includes only elements that affect the A-UGV performance. An example environment standard is ASTM F3218 [14]. Based on the described factory environment, users can identify requirements of A-UGVs needed. Lastly, there needs to be a common understanding of A-UGV performance and the factory environment between makers and users, which these standards should provide. ASTM WK65139 A-UGV Capabilities standard includes both A-UGV performance and environmental conditions in which the A-UGV can prove or assert capabilities. The next section describes how standards are used and support common understanding.

2.2 How standards are used

A-UGV users can check the manufacturing environments defined in the ASTM F3218 standard and whether they exist in their manufacturing system. By measuring the vehicle workspace, including all possible conditions that could affect the A-UGV performance, the environment can be replicated. For example, the user can not only define narrow paths and turns but also potential challenges (e.g., low lighting, floor defects) with their exact locations noted from, for example, an intersection


Standards for
A-UGV performance test
Test framework
Defined Area
L-Shape Path
Environmental Condition
Lighting
Floor Conditions
A-UGV Maker
1. See what should be tested & and how to test
2. Replicate the elements & Test performance
3. Develop Standard based documents
A-UGV user (factory owner)
1. See what factory element should be defined
2. Define elements existing in the factory
3. Develop Standard based documents
A-UGV Measured Performance Information
Specified Factory Information
Start Point
40 Lux
2 m
6 m
Light source
350 Lux
End Point
2 m
Turn
5 m
<Replicated test area>
<Existing factory elements>

(Fig. 2 right). Then, users can ask makers about the expected A-UGVs performance at user's factory by providing detailed operating environment information.

An A-UGV maker can build a test system to evaluate the performance of A-UGV for items defined in the standard. For example, in order to evaluate driving performance on narrow paths and turns, they can measure and evaluate the performance of an A-UGV by establishing the narrowest path environment in which their A-UGV can travel, as shown in Fig. 2 left. In addition, it is possible to measure the obstacle avoidance performance by making obstacle models and installing them in the driving path.

With the process shown above, the user and maker can communicate efficiently and effectively. The user can provide more exact A-UGV operating environment information and the maker can provide A-UGV performance values for the relevant environment. The user can find the A-UGV model that best suits their manufacturing environment and the maker can sell the A-UGV to users who have environments where their A-UGVs can operate.

In addition to helping to communicate between the maker and the user, standards help to clarify what to discuss. The standards provide the list of items that may affect A-UGV performance. The next section discusses items to be defined as standards.

2.3 Standards items to be defined

There are numerous factors that affect the A-UGV performance and many performance indexes. According to the purpose of this study, the standard items include only factors that: 1) can describe the operating environment, 2) can affect A-UGV functionality, especially navigation and docking, and 3) can commonly exist in factories. Items in ASTM F45 standards include the A-UGV performance descriptions, tests, and measurements that: 1) describe functionality against the factors, 2) test and document navigation, docking, and/or other vehicle tasks and capabilities, and 3) measure the A-UGV performance regardless of technology (e.g., automatic or autonomous). There are two types of factors, referred to as environmental factors and obstacles.

Environmental factors refer to the environmental conditions in which the A-UGV operates. If an A-UGV navigates successfully under an environmental factor, the A-UGV is determined to provide the expected performance for the factor. Environments can be classified into floor factors such as: ramps, gaps, gouges, and steps, and non-floor factors such as: lighting, temperature, humidity, and blocks.

ASTM F3200 defines obstacle as: "static or moving object that obstructs the intended movement" [15]. Obstacles in the vehicle path must be avoided (e.g., stop or navigate around) to not cause collisions or other unexpected A-UGV performance. To avoid obstacles, automatic A-UGVs stop, whereas autonomous A-UGVs can have the ability to recognize, measure the size of, and regenerate path(s) to avoid obstacles. There are various kinds of obstacles in factories depending on target products and processes of manufacturing systems. Since the A-UGV performance is affected by the obstacle characteristics (e.g., size, shape, reflectivity), it is necessary to analyze obstacles in with regards to how they affect A-UGV performance.

A-UGV performance refers to, among other capabilities, what the vehicle measures and reacts to in the presence of

TABLE 1: ENVIRONMENT AND OBSTACLE FACTORS, THEIR IMPACT TO THE A-UGV, AND THE REQUIRED A-UGV FUNCTIONS

Factors	Impact to A-UGV	Required A-UGV functions
Environment		
narrow curved path	stuck at the corner	to control minimum clearance from boundaries
grade (ramp)	obstacles by floor level change negative obstacles by floor level changes stuck at the entrance stuck before exit navigation error (vehicle location lost)	to recognize presence of a ramp to recognize being on a ramp to transit from floor to a ramp to recognize being stuck to recognize location lost
gap, step	stuck by gap or step floor detected as obstacle by tilted vehicle	to pass a gap, to recognize a gap or step and avoid to recognize being tilted
floor materials	docking / undocking failure by wheel slip	to drive on a surface without slip
transparent curtain barrier	blocked by barrier collide with obstacles near barrier	to pass barrier to detect obstacles near barrier
Obstacle		
general obstacle	fail to navigate due to collision damage to vehicle from collision	to detect obstacles before collision to measure the size and location of obstacles to stop before collision to bypass obstacles
small obstacle	fail to navigate and damage from collision	to detect obstacles near floor
overhanging obstacle	fail to navigate and damage from collision	to detect obstacles above floor below A-UGV height
moving obstacle	fail to navigate and damage from collision	to control time to reserve detected obstacle to detect obstacles frequently to predict next move of detected obstacle
lighting obstacle	stuck in front of lighting obstacle	to detect objects against direct light
transparent obstacle	fail to navigate and damage from collision	to detect transparent obstacle
negative obstacle	fail to navigate from fall severe damage to vehicle from fall	to detect negative obstacle to determine feasibility to pass negative obstacle
virtual obstacle	invade to human work or walk spaces	to behave against virtual obstacle as general obstacle

environmental factors and obstacles. The simplest navigation test result is pass or fail and in many failure cases, it is worth noting why it failed. For example, when an A-UGV is navigating a narrow path with an obstacle, the A-UGV performance can be evaluated differently. Failure examples could include: fails to measure the size of the obstacle, fails to determine a bypass route, fails to regenerate a route, approached and became trapped between the obstacle and the wall, or failed to pass narrow path. Alternatively, the vehicle may have performed as expected and simply stopped and alerted a supervisor of the issues.

Standards that document A-UGV performance should allow test requestors to clearly and concisely describe under which condition the vehicle is expected to pass against factors. Thus, standards should expect the test requestor to clearly document minimum and maximum values of measurable factors. For example, an A-UGV may be expected to successfully drive over 0 mm to 5 mm maximum steps.

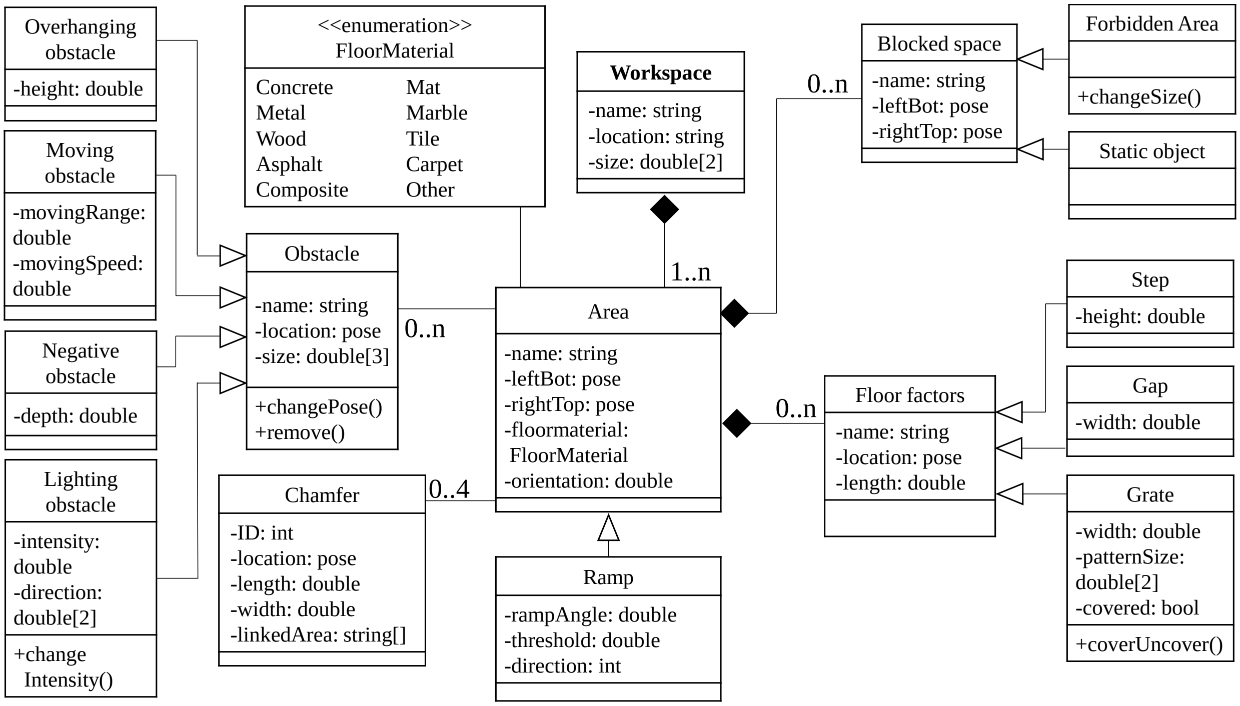
Through previous studies [7], various environment and obstacle factors affecting A-UGV performance were identified and their characteristics were analyzed. Through subsequent research, several factors affecting A-UGV performance and the functions required for the A-UGV to respond are summarized in Table 1.

For each factor listed in the table, impact to A-UGV means there is a possibility that it may occur, and not all A-UGVs may be affected. Required functions for A-UGV means the capability is needed to respond to the factors. However, current F45 performance standards only provide test methods to measure the A-UGV performance regardless of how the vehicle technology is designed and handles factors. The intent in F45 standards is to allow the test requestor flexibility to determine how their A-UGV should handle each factor. For example, to solve a the situation when an A-UGV is stopped by a step, any solution making the vehicle surpass the step can be introduced. And, through standard tests requested, it is to be proven that the A-UGV can in fact do so.

In this paper, the effect of A-UGVs responding to a single factor, i.e., environment or obstacle, is described. However, more complex cases exist at factories where both simultaneously exist. Through various experiments, it was confirmed that the combination of both environmental factors and obstacles had additional effects to lower A-UGV performance. The method for defining such a complex situation and evaluating the corresponding A-UGV performance will be carried out through subsequent studies.

3. INFORMATION MODEL

A-UGV performance factors should be addressed in terms of target factories and A-UGVs rather than individually. In other words, each factor should be expressed as a property of the



factory or the A-UGV. At the same time, it must be able to express the factory elements (defined here as factory environment and obstacle(s)) independently, and relationships between the elements should also be revealed. The same is true of the A-UGV performance.

Accordingly, this section develops information models to support development of A-UGV performance measurement standards. The information model lists the elements necessary for the performance evaluation of A-UGVs and expresses how each element is related. The information model follows the class diagram of UML.

In this model, a class represents independent factory elements or the performance of an A-UGV. For example, obstacles in a factory can be defined as one class. Each class can have attributes describing the class. For example, obstacle class has location and size attributes. Each class is connected to other classes with relationship(s). For example, obstacle class is connected to area class, which means an obstacle exists in a factory area. Each class can define actual factory elements (see Fig. 6) or A-UGV performance (see Fig. 7) by specifying attribute values, called instances or objects. For example, a safety cone can be an obstacle class object that embodies size and location information.

An information model representing both the factory and A-UGV performance was developed to focus on classes to be defined and their relationships. Several attributes associated with factory elements can be added for each class, but those only suitable for this paper (i.e., considering only environmental and obstacle factors) are listed based on Table 1.

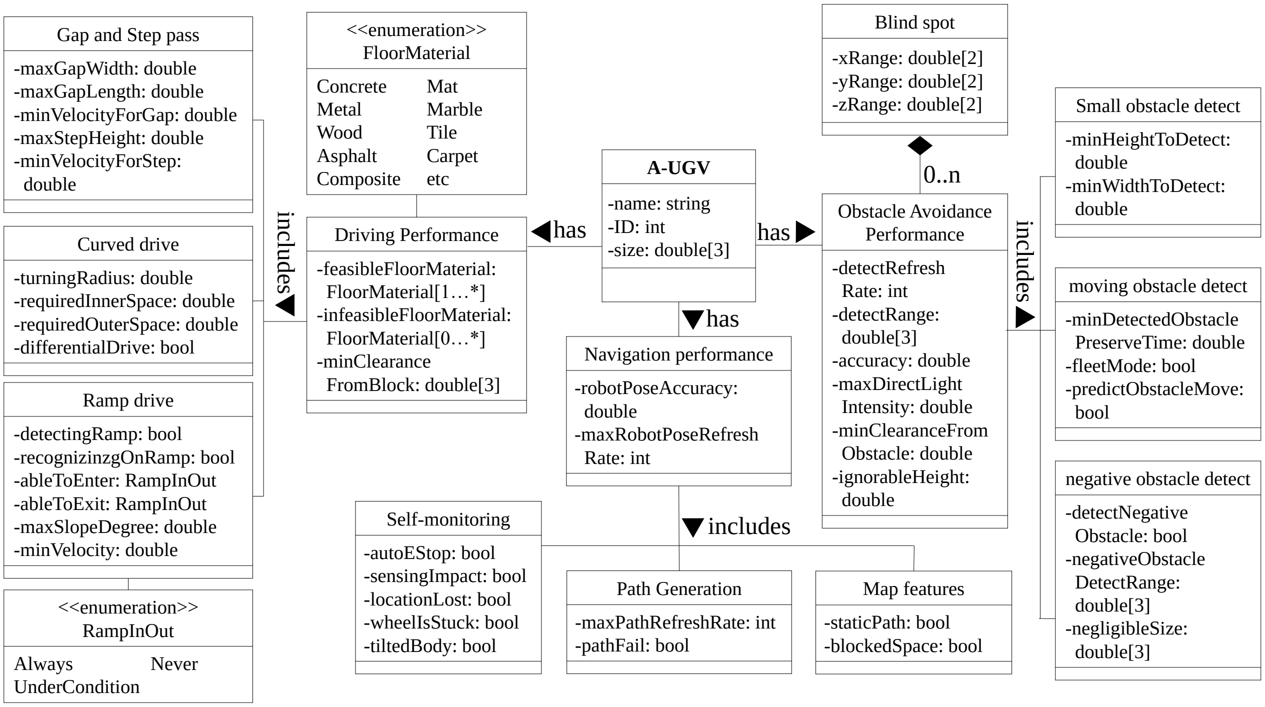
3.1 Workspace information model

In terms of A-UGV introduction, the most important part of a factory is navigation areas. From the perspective of the A-UGV, an area can be referred to as a confined space where A-UGVs can drive. Other spaces independent of A-UGVs are not included in the model even though they may play an important role in manufacturing processes. This paper uses the term ‘workspace’ as where the A-UGV can automatically navigate as specified by the user. The workspace can consist of one or more areas and the A-UGV can move freely between areas within the workspace. Therefore, the factory information model needs to be built in units of workspaces. Figure 3 shows the workspace information model.

Workspace is defined as base class. It has name, location, and size as attributes. Location is used to describe where a workspace is in a factory and workspace must have at least one area.

Areas are defined as squares by default. The location and size of the area are described by the left-bottom and right-top coordinates in workspace. It is necessary to specify the floor material. Common floor materials are concrete, metal, wood, asphalt, mat, marble, tile, and carpet. Orientation is used when the reference origin of the area and workspace are different.

A ramp area is defined as a derivative of the area. A-UGV performance may be affected by a ramp, especially according to angle and ramp transition. The ramp information model includes direction properties to express incline or decline.



An area can have blocked space where the A-UGV should not access. Blocked space can be specified as a forbidden area or static object. A forbidden area is typically to prevent A-UGV collision with other objects, such as a pedestrian walkway. Static objects refer to those that are difficult to detect and rarely move, such as a workbench or a steel beam stacked on a shelf.

An area may have floor factors including irregular floor level changes. Gap, step, and grate are the typical types of floor factors. As floor factors may disturb and stop A-UGVs, their locations must be described. The size of floor factors should also be specified including height or depth information. Grates can be described as a set of patterned gaps. The pattern size describes the property of a grate. A grate can be changed to a negative obstacle when it is removed.

An area may have chamfered corners. The chamfer information model describes where it is, and to which area it is linked. The shape of a chamfer is described by length and width. Detailed measurement and expression methods are defined in the ASTM F3244 navigation standard [13].

An area may also have various obstacles. Each obstacle contains the location in the area and size information. Obstacles can be moved or removed. There are derivatives of the obstacle as the A-UGV may require further information to avoid them. They have additional attributes that may affect A-UGV performance.

3.2 A-UGV information model

The A-UGV information model begins by defining a target A-UGV class (Fig. 4). It includes basic information such as name, identification, and size. The A-UGV class has driving performance, obstacle avoidance performance, and navigation performance classes, and each of them have subclasses. Each class describes performance in each domain to respond to the factors defined in the workspace information model.

Driving performance class describes the environment in which the A-UGV can drive. Floor materials in areas where the A-UGV can or cannot drive should be specified. The class includes the minimum clearance from blocks required for safe driving. Driving performance includes gap and step pass, curved drive, and ramp drive classes.

Gap and step pass class describes maximum gap width and length, and maximum step height that the A-UGV can pass, and minimum velocities needed to pass the gap and step.

Curved drive class describes the required inner and outer radii in which the A-UGV can drive. The class also includes whether or not the A-UGV has differential steering which may allow the A-UGV to navigate tighter turns.

Ramp drive class describes abilities to recognize the presence of a ramp and that the vehicle is on a ramp. In addition, there are attributes to describe capabilities to drive on and off ramps. The class has attributes about the maximum ramp angle and minimum velocity required to climb the maximum ramp angle.

Obstacle avoidance performance class describes the capabilities for the A-UGV to detect and avoid obstacles. The

Yoon, Soocheol; Bostelman, Roger V. "Information Model For A-Ugv Performance Measurement Standard." Presented at International Mechanical Engineering Congress and Exposition (IMECE 2020). November 16, 2020 - November 19, 2020.

ability to detect obstacles includes refresh rate, range, and accuracy. It includes lighting conditions in which sensors can operate normally. The ability to avoid obstacles includes minimum A-UGV clearance required to bypass obstacles and a minimum height value to ignore for overhanging obstacles under which the A-UGV can drive.

There are small, moving, and negative obstacles as subclasses of obstacle avoidance performance class. The small obstacle class describes the minimum height and width obstacle that the A-UGV can recognize. In the moving obstacle class, a minimum time to preserve detected objects in a navigation map is defined to avoid moving obstacles safely. A fleet mode support is defined to integrate other vehicles in a workspace. Predict function support is defined to predict next moves of moving obstacles. Negative obstacle class describes whether the A-UGV can recognize negative obstacles and the recognizable ranges. A-UGVs can ignore and pass over negative obstacles when one obstacle's depth, width, or length is small enough. Negligible sizes attribute describes their maximum values.

Depending on the type and arrangement of the sensors, blind spots exist where objects may be undetected. This usually occurs in spaces between A-UGV sensors. Blind spot class describes poses of A-UGV sensing volumes. An A-UGV may have no or multiple blind spots.

Navigation performance class is composed of the capability to monitor the current status and respond. Navigating begins with knowing the current pose of the A-UGV. Pose accuracy and refresh rate attributes are defined for the pose performance. Navigation performance class has self-monitoring, path generation, and map features subclasses.

Self-monitoring class defines several capabilities to recognize current status of the vehicle body and when an automatic emergency stop can occur. A-UGVs may be able to recognize physical impacts to the body, lost vehicle pose from the navigation map, drive wheels stuck from gaps, steps, or other floor factors, and vehicle tilt by a ramp, step, or uneven floor. Path generation class describes how frequently paths are updated and the capability to determine infeasible paths. Map feature class describes the capability to place paths and forbidden areas in the navigation map for efficient and safe A-UGV navigation.

4. CASE STUDY

As a case study, object diagrams were developed as information model instances for a National Institute of Standards and Technology (NIST) testbed having a similar environment to a factory. The testbed contains a lot of the spaces and objects including ramps, grate, steps, workshops, desks, chairs, machines, storage spaces, safety cones, etc. The main purpose of the testbed, in this case study, was to test A-UGV performance on and around ramps. It includes the abilities to undock from various points, to navigate to ramp entrances, to climb ramps, and to achieve a goal positioned at the level ramp top. The testbed has two ramps with 5˚ and 10˚ slopes. Figure 5 top shows the top view of the testbed and outlines the defined target workspace. Figure 5 bottom describes areas needed to be modeled. There are three steps and a grate as floor factors, and tires (T1 and T2), trash bins (B1 and B2), and safety cone (C) as obstacles.

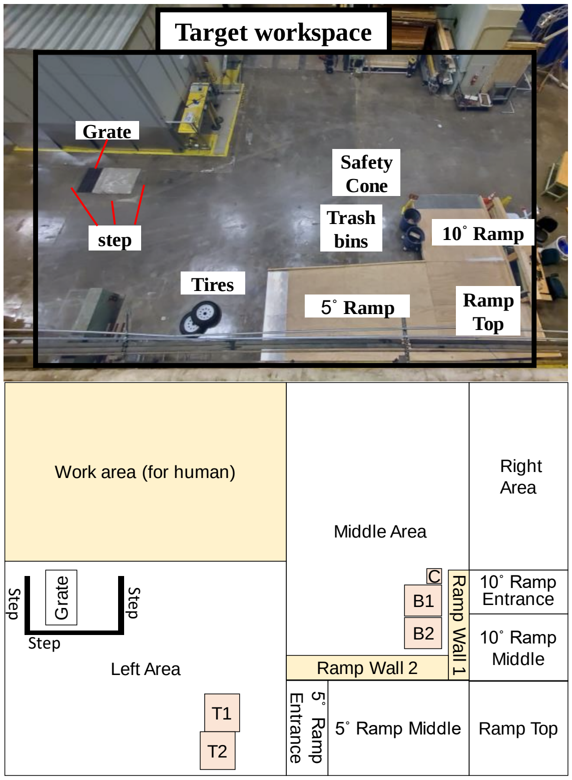


FIGURE 5: THE TARGET WORKSPACE PICTURE (TOP) AND DIAGRAM (BOTTOM)

Using the workspace information model defined in this paper, the instance for the target workspace is shown in Fig. 6 object diagram. A workspace object is defined as base with location and size information.

Four area and four ramp objects are defined under the workspace with name, location, and floor materials. Ramp structure consists of entrance, middle, and top objects. Entrances have 7 mm and 5 mm steps each, and a metal floor. Ramp middle and top have wooden floors. The direction attributes describe the direction of the ramp incline.

Left area object has forbidden area, obstacles, steps (i.e., in this case, small raised area) and grate objects. Forbidden area describes the location of the human work area. The Obstacle objects describe location and size of the folk lift and tires. The step and grate objects describe their locations and sizes.

Middle area object has obstacle and static object objects. Obstacle objects describe the location and size of trash bins and the safety cone. Static object describes the side of ramps, namely ramp walls, as they need to stop the vehicle when approaching the ramp from the wrong direction.

An A-UGV had been tested to measure its performance against the factors defined in the workspace object diagram. An object diagram is developed for A-UGV performance as shown

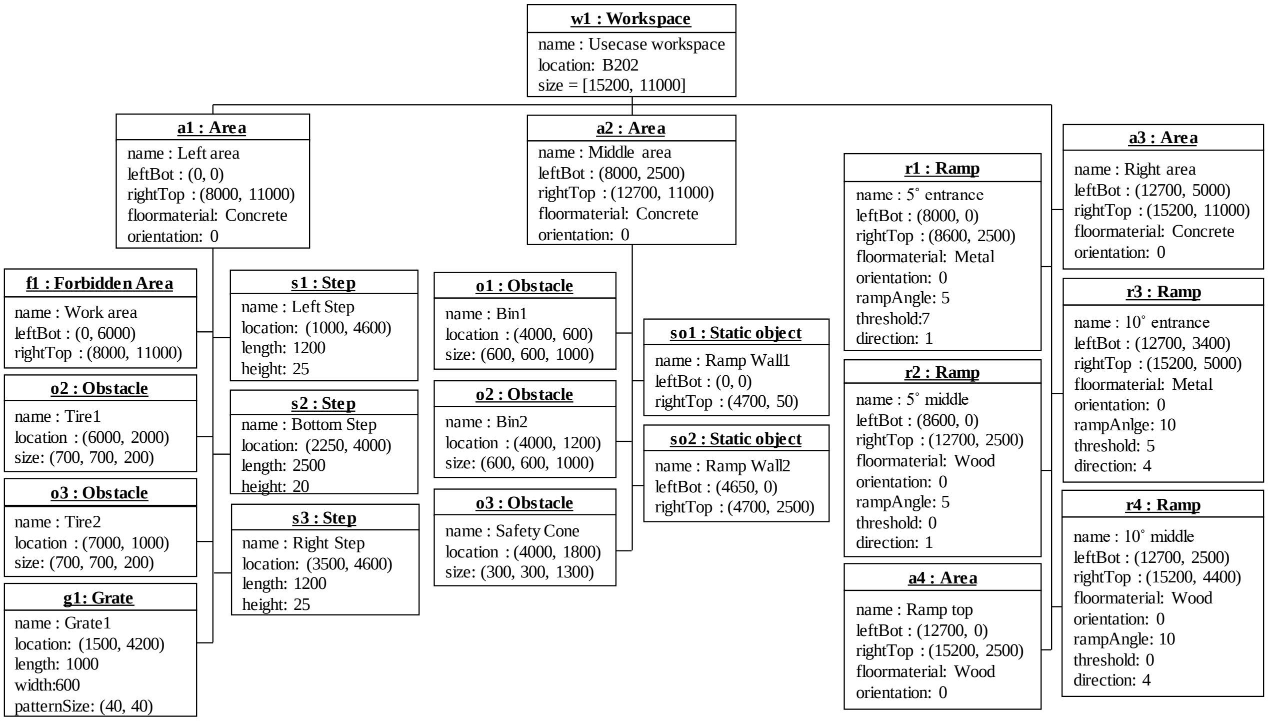


FIGURE 6: THE TARGET WORKSPACE OBJECT DIAGRAM

in Fig. 7. An A-UGV object is defined as base with name and size information. It has driving, navigation, and obstacle avoidance performance objects.

The driving performance object describes the performance against the floor factors. The A-UGV can drive on concrete, wood, and metal floor. The A-UGV required 50 mm clearance from blocks and succeeded to pass over a 20 mm height step at 500 mm/s velocity. When the A-UGV turned 180˚, it required 80 mm inside diameter clearance. The A-UGV could neither detect the ramp nor recognize being on the ramp. The A-UGV succeeded to enter and exit the ramp only when the side sensors were off. The A-UGV succeeded to climb 10˚ ramp with 300 mm/s speed.

The navigation performance object describes the performance of how the A-UGV perceives its status. The A-UGV had measured its pose under 100 mm error with 15 Hz refresh rate. The A-UGV stopped automatically when it failed to enter the ramp. The A-UGV could neither recognize that the drive wheels were stuck nor the tilted vehicle body by the steps. The A-UGV had set blocked spaces at the location of the ramp walls.

The obstacle avoidance performance object describes the performance against obstacles. The A-UGV had detected obstacles in the range of 8000 mm x 8000 mm x 1300 mm with 15 Hz frequency. The A-UGV detected obstacles correctly within 20 mm error and required 50 mm clearance to avoid obstacles. The A-UGV was able to detect obstacles equal to or larger than 200 mm x 300 mm. The A-UGV failed to detect the holes in grate, which are 40 mm x 40 mm x 1000 mm but the A-UGV could drive over the grate.

Through this case study, it is shown that the workspace information model can describe factors affecting A-UGV performance easily and efficiently. Also, it is shown that the A-UGV performance information model can describe its performance against those factors.

5. CONCLUSION

This paper analyzed needs of A-UGV performance standards and developed an information model for A-UGV performance. The information model supports A-UGV makers and users to describe and discuss the factors that affect A-UGV performance using a standard model, and the performance against them by providing the workspace model and A-UGV performance model. The workspace information model describes areas where A-UGV may drive with floor factors and obstacles information. The A-UGV information model describes the abilities to drive, navigate, and avoid obstacles responding to the factors defined in the workspace model. The case study was performed by applying the information model to the testbed having a mock factory environment. Through the case study, it was shown that the information model can describe the factory and A-UGV easily and efficiently in structured form for performance evaluation purposes.

This paper contributes to 1) help mutual understanding between A-UGV manufacturers and users through A-UGV and factory information models, 2) enable quick and structured

Yoon, Soocheol; Bostelman, Roger V. "Information Model For A-Ugv Performance Measurement Standard." Presented at International Mechanical Engineering Congress and Exposition (IMECE 2020). November 16, 2020 - November 19, 2020.

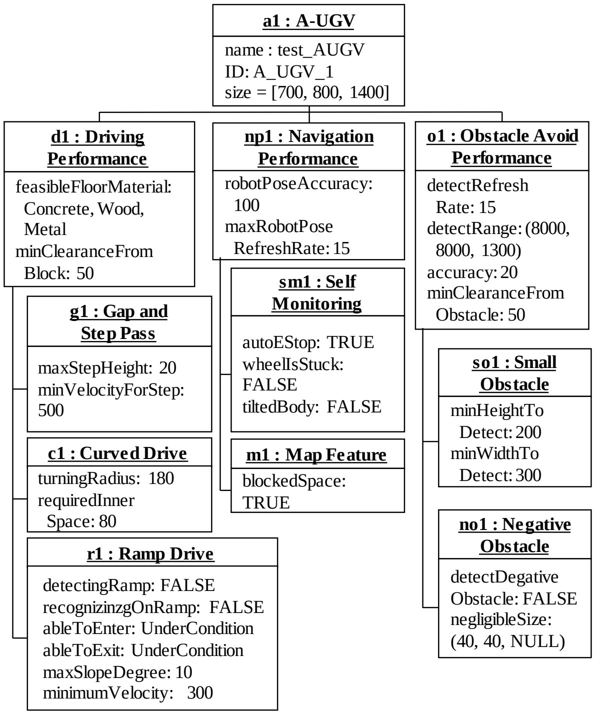


FIGURE 7: THE A-UGV PERFORMANCE OBJECT DIAGRAM

information delivery, and 3) provide a basis for testing various A-UGVs under the same environment in various locations.

A-UGV performance evaluation methods for more complex situations are planned to be conducted. This includes a combination of various floor conditions and obstacles and in a situation where multiple obstacles with different characteristics exist, such as transparent and luminous objects. Also, heterogeneous vehicles will be tested simultaneously in the same workspaces to see how the performance would change when they are grouped. The target and testing manufacturing system will be extended to larger scale with framework to manage the A-UGV in both hardware and software. The results of each study will be reflected in the information model and recommended to standards bodies.

REFERENCES

[1] Hamner, Brad, Seth Koterba, Jane Shi, Reid Simmons, and Sanjiv Singh. "An autonomous mobile manipulator for assembly tasks." *Autonomous Robots* Vol.28, No.1 (2010): pp.131-149. DOI 10.1007/s10514-009-9142-y

[2] Ren, Shunan, Ying Xie, Xiangdong Yang, Jing Xu, Guolei Wang, and Ken Chen. "A method for optimizing the base position of mobile painting manipulators." *IEEE Transactions on Automation Science and Engineering* Vol.14, No.1 (2016): pp.370-375. DOI 10.1109/TASE.2016.2612694

[3] ASTM F45, Committee F45 on Driverless Automatic Guided Industrial Vehicles, https://www.astm.org/COMMITTEE/F45.htm

[4] Mehami, Jasprabhjit, Mauludin Nawi, and Ray Y. Zhong. "Smart automated guided vehicles for manufacturing in the context of Industry 4.0." *Procedia Manufacturing* Vol.26 (2018): pp.1077-1086. DOI 10.1016/j.promfg.2018.07.144

[5] Weyer, Stephan, Mathias Schmitt, Moritz Ohmer, and Dominic Gorecky. "Towards Industry 4.0-Standardization as the crucial challenge for highly modular, multi-vendor production systems." *Ifac-Papersonline* Vol.48, No.3 (2015): pp.579-584. DOI 10.1016/j.ifacol.2015.06.143

[6] Bostelman, Roger, Sebti Foufou, Tsai Hong, and Mili Shah. "Model of Mobile Manipulator Performance Measurement using SysML." *Journal of intelligent & robotic systems* Vol.92, No.1 (2018): pp.65-83.

[7] Yoon, Soocheol, and Roger Bostelman. "Analysis of Automatic through Autonomous-Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards." *In 2019 IEEE International Symposium on Robotic and Sensors Environments* (ROSE), pp.1-7. Ottawa, ON, June 17-18, 2019. DOI 10.1109/ROSE.2019.8790421

[8] Bostelman, Roger, Tsai Hong, and Jeremy Marvel. "Survey of research for performance measurement of mobile manipulators." *Journal of Research of the National Institute of Standards and Technology* Vol.121, No.3 (2016): pp.342-366. DOI 10.6028/jres.121.015

[9] Young, Stuart H., Thomas A. Mazzuchi, and Shahram Sarkani. "A framework for predicting future system performance in autonomous unmanned ground vehicles." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* Vol.47, No.7 (2016): pp.1192-1206. DOI 10.1109/TSMC.2016.2563403

[10] Wu, Bichen, Forrest Iandola, Peter H. Jin, and Kurt Keutzer. "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.129-137. 2017.

[11] ISO, ISO/IEC 19051:2005 Information Technology – Open Distributed Processing – Unified Modeling Language (UML) Version 1.4.2

[12] Lu, Yan, Katherine C. Morris, and Simon Frechette. "Current standards landscape for smart manufacturing systems." *Technical Report No.NISTIR 8107*, National Institute of Standards and Technology, Gaithersburg, MD. 2016. DOI 10.6028/NIST.IR.8107

[13] ASTM International, *ASTM F3244-17 Standard Test Method for Navigation: Defined Area*, West Conshohocken, PA; ASTM International, 2017. DOI: https://doi.org/10.1520/F3244-17

[14] ASTM International. *F3218-19 Standard Practice for Documenting Environmental Conditions for Utilization with A-UGV Test Methods*. West Conshohocken, PA; ASTM International, 2019. doi: https://doi.org/10.1520/F3218-19

[15] ASTM International. *F3200-19 Standard Terminology for Driverless Automatic Guided Industrial Vehicles*. West Conshohocken, PA; ASTM International, 2019. doi: https://doi.org/10.1520/F3200-19

FREQUENCY TUNABLE LABEL-FREE SURFACE ACOUSTIC WAVE-BASED FLOW SENSOR

Aurore Quelennec, Jason J. Gorman, and Darwin R. Reyes*
National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

ABSTRACT

We present a label-free surface-acoustic wave (SAW)-based flow sensor with enhanced precision and repeatability. This sensor improves the signal-to-noise ratio by one order of magnitude (dB scale) at an optimized frequency, and lowers the measurable flow rate to 3 µL.min^{-1}, from 250 µL.min^{-1}, when compared to our previous approach using a sine wave generator. The device was designed so that the distance between the transducers is tunable to further increase its sensitivity to 53 mdB.µL^{-1}.min. This easy-to-integrate sensor does not affect the microfluidic network and offers an integrated flow measurement.

KEYWORDS: Flow Metrology, Electroacoustics, SAW Sensor

INTRODUCTION

Sensing capabilities of SAW-based sensors include flow measurements, which have previously been demonstrated in microfluidic systems [1]. Such SAW flow sensors based on thermal, doppler shift or time-of-flight measurements require labeling of the fluid. Our label-free SAW-based flow sensor relies on absorption of the SAW by the liquid as a function of the flow rate and SAW's frequency. This work presents label-free flow measurements over the range of 0 µL.min^{-1} to 1 mL.min^{-1} in a microfluidic channel. The design consists of a microchannel that passes between the transducers, which are placed in air cavities to improve the accuracy of the measured flow.

EXPERIMENTAL

The sensor is fabricated on a 128XY (128º X-Y cut) lithium niobate substrate. It has two pairs of interdigitated transducers (IDT), one emitter and one receiver, placed in air cavities on either side of the microchannel (Figure 1). Each transducer has 85 pairs of fingers made out of titanium (10 nm) and gold (90 nm). All fingers are 0.5 mm long and have a pitch (λ) of 80 µm. A 100 nm thick silicon oxide layer is sputtered as an adhesion layer for the 80 µm deep polydimethylsiloxane (PDMS) microchannel. We fabricated three pairs of IDT, each with a different distance traveled by the SAW in the liquid (*Df*): 0.7 mm, 0.9 mm and 1.1 mm. The deionized water flow is controlled by a syringe pump at rates between 0 µL.min^{-1} to 1 mL.min^{-1}, in steps of 50 µL.min^{-1} (± 0.35 %, as certified by the manufacturer). Each measurement is done in triplicate. The two IDTs are connected to a vector network analyzer (VNA) to measure the transmission coefficient (S12) at each flow rate for frequencies between 45 MHz and 46 MHz, with a sampling frequency of 1.88 kHz. Contrary to the sine wave generator used in our previous work [2], the VNA is a highly-accurate measurement device, with an uncertainty of less than 0.2 dB for an S12 measurement result between -70 dB and -50 dB, and an uncertainty less than 0.04 dB for an S12 measurement result above -50 dB.

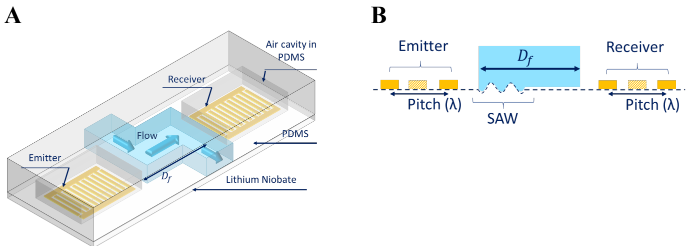


Figure 1: Sensor configuration. A) Sketch of one sensor. Emitter and receiver are placed on a lithium niobate substrate and on opposite sides of the microchannel where the liquid flows. The microchannel wall is in PDMS. B) Cross-section sketch. The SAW travels a distance Df at the liquid-solid interface, from the emitter to the receiver. The IDT's pitch is 80 µm.

RESULTS AND DISCUSSION

The acoustic absorption of a medium is dependent on the frequency of the SAW and *Df* [3]. S12 amplitude depends on *Df* (compare Figure 2A, B, C, or 2G, H, I at no flow) and frequency (Figure 2E, F, G). The maximum amplitude of S12 is reached at the acoustic frequency and decreases with the increase in *Df*, as expected [3]. The highest sensitivity to flow of 53 mdB.µL^{-1}.min is obtained at the low-notch frequency with *Df* = 0.9 mm (Figure 2H). With *Df* = 0.9 mm and a flow range between 0 µL.min^{-1} and 450 µL.min^{-1}, we obtained a precision of 0.2 dB and a

Quelennec, Aurore; Gorman, Jason J.; Reyes-Hernandez, Darwin. "Frequency Tunable Label-Free Surface Acoustic Wave-Based Flow Sensor." Presented at 24th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MIcroTAS 2020). October 04, 2020 - October 09, 2020.

theoretical limit of detection of 4 µL/min. As the flow rate increases, S12 amplitude decreases and then increases. Thus, a slip is likely responsible for the non-monotonic response of the S12 to flow rate (Figure 2G, H). In fact, McHale *et al.* [4] have previously reported that the absorption is optimal when the slip can be neglected, and starts decreasing when the liquid slips on the solid. The best precision of 0.04 dB is obtained at the notch frequencies with *Df* = 1.1 mm (Figure 2I), given a theoretical limit of detection of 3 µL.min^{-1} with a sensitivity of 16 mdB.µL^{-1}.min.

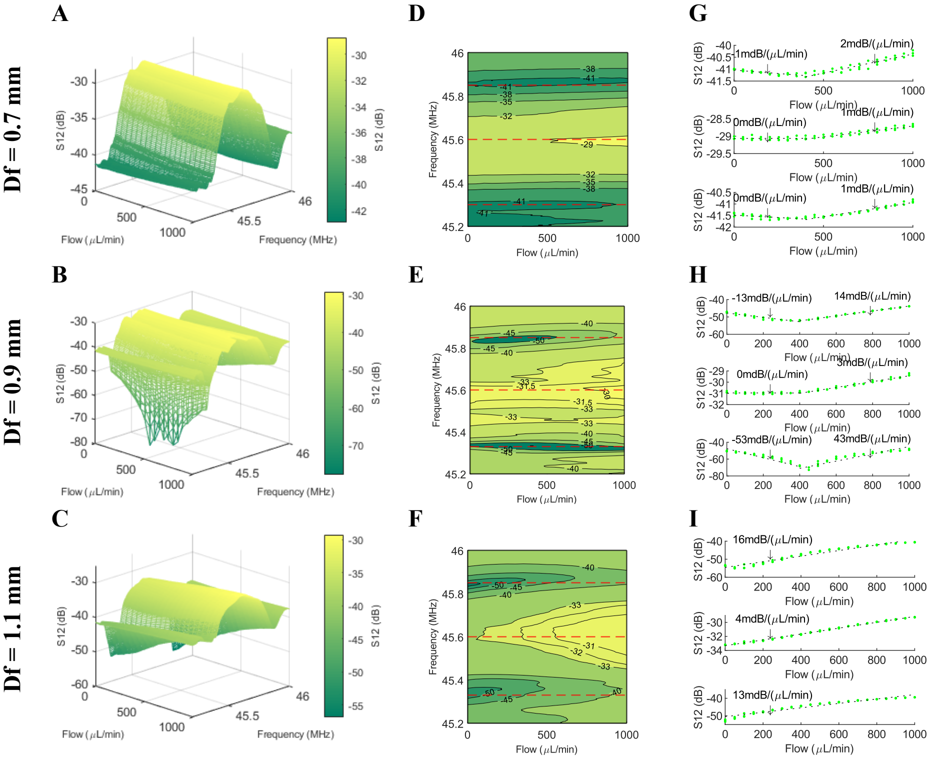


Figure 2: Sensors' responses to flow for Df of 0.7 mm, 0.9 mm and 1.1 mm. A-B-C) 3D plot of S12 as a function of the frequency and the flow rate. D-E-F) Colormap of S12 as a function of the frequency and the flow rate. Three horizontal dashed red lines are plotted at the low-notch frequency, the acoustic frequency and the high-notch frequency, from bottom to top respectively. S12 is the most sensitive to flow rate at these three frequencies. G-H-I) S12 of three measurements (green) and their fit (dashed line) as a function of the flow rate and at the low-notch frequency, the acoustic frequency and the high-notch frequency (from bottom to top). Df and frequency can be tuned to increase the sensitivity to the flow rate. The sensitivity is the highest at the low-notch frequency for Df = 0.9 mm. A wider range of flow rate is measurable, using Df = 1.1 mm.

CONCLUSIONS

This work presents a label-free SAW-based flow sensor with greatly improved accuracy that can be integrated in a microfluidic chip to provide local flow data. This sensor could be integrated, in the future, in drug encapsulation droplet and in diffusion-based reaction microfluidic systems where localized flow measurements are critical to produce high yield results.

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] J.T.W. Kuo, L. Yu and E. Meng, *Micromachines* 2012, 3(3), 550-573; doi:10.3390/mi3030550
[2] A. Quelennec, J.J. Gorman and D.R. Reyes, in *Proceeding of MicroTAS 2019*, Basel, Switzerland, 2019
[3] Arzt, R. M., Salzmann, E., and Dransfeld, K. (1967). *Appl Phys Lett*, 10(5), 165–167. doi:10.1063/1.1754894
[4] McHale, G., *et al.* (2000). *Mat Sci Eng C-Mater*, 12(1-2), 17–22. doi:10.1016/s0928-4931(00)00151-x

CONTACT: Darwin R. Reyes; phone: +1-301-975-5466; darwin.reyes@nist.gov

FPGA Implementation of a Low Latency and High SFDR Direct Digital Synthesizer for Resource-Efficient Quantum-Enhanced Communication*

N. Fajar R. Annafianto*[,1,2], M.V. Jabir[2], I.A. Burenkov[2], H.F. Ugurdag[1], A. Battou[2] and S.V. Polyakov[2]
[1]Electrical and Electronics Engineering Dept., Ozyegin University, Istanbul, Turkey
[2]National Institute of Standards and Technology, Maryland, USA
*annafianto.annafianto@ozu.edu.tr

***Abstract*—A Direct Digital Synthesizer (DDS) generates a sinusoidal signal, which is a significant component of many communication systems using modulation schemes. A CORDIC algorithm offers minimum memory requirements compared to look-up-based methods and low latency. The latency depends on the number of iterations, which is determined by the number of angles in the rotation set. However, it is necessary to maintain high spectral purity to optimize the overall system performance. To optimize the opportunity of quantum measurement, low latency and a high spectral purity sine wave generator is essential. The implementation of this design generates output with 64% latency reduction compared to that of the conventional CORDIC design and 72.2 dB SFDR value.**

Keywords—FPGA, CORDIC, DDS, SFDR, pipeline, latency.

I. Introduction

In most modulation schemes for a digital telecommunication system, a fast and efficient sinusoidal signal generator is needed. Here we report on an FPGA implementation of a versatile Coordinate Rotation Digital Computer (CORDIC) based Direct Digital Synthesizer (DDS). Most commercial lightwave communication systems use standard modulation protocols, such as Phase-Shift Keying (PSK) and Frequency-Shift Keying (FSK), whose implementation is supported by specialized dedicated hardware. There is a need for significant improvement in energy and bandwidth efficiency. Therefore, to gain further improvement a new class of communication systems, namely quantum-measurement enhanced optical communication systems are being actively pursued. In those systems, a classical receiver is replaced with a quantum receiver, while the transmitter remains the same. Properties of quantum measurement are in general different from that of classical measurement. They require more complex modulation schemes than PSK and FSK [1]. Digital synthesis of these signals requires versatile DDS whose development is reported in this work. By design, a DDS generates signals with a nearly arbitrary combination of phase and frequency modulations. Many other applications such as software-defined radio, wireless satellite transceiver, HDTV transmission, radar communication, etc. can take advantage of this low latency and re-configurable sine wave generation [2]. With its high spectral purity and low latency, the DDS accommodates the energy-efficient and rapid response properties needed by quantum measurement instruments in order to surpass efficient of classical measurement and maximize the modulation capabilities.

Many strategies and techniques have been developed to enhance the hardware area and speed efficiency of CORDIC algorithm implementation. CORDIC was initially introduced by Volder in 1959 to calculate trigonometric functions in digital hardware devices [2]. Later, a modified version of a CORDIC algorithm was proposed by S. Walther with the ability to calculate circular, hyperbolic, and linear rotation systems [4]. The motivation to use this algorithm in digital platforms has gained popularity since then. Refinements on the efficiency of implementation have resulted in reduced latency and hardware area usage.

In the next section, we provide background information on several CORDIC techniques that are adopted in this work. In section III, we explain the implementation of the proposed method. In section IV, we summarize and compare the obtained results. Finally, we conclude with an evaluation and discussion of the prospective developments.

II. Background

The demand for higher-throughput communication has always existed and provides the environment for developing new applications. Enhancing the performance of a DDS will lead to a throughput increase. Many communication systems use a modulation scheme that generates sinusoidal signal output. One popular approach is to use a DDS hardware module that takes the Frequency Tuning Word (FTW) or Frequency Control Word (FCW) as input and passes the amplitude of the sinusoidal signal to the output. Figure 1 shows the general block diagram of a DDS which consists of a phase accumulator, a signal processor, Digital-to-Analog Converter (DAC) and a low pass filter. The phase accumulator defines the frequency of the sinusoidal signal as the increment of the output phase that is dictated by the input FCW, hence the smaller the input the lower the resulting signal's frequency and vice versa. The low pass filter removes aliasing of the signal processor's output that contains some noise due to the techniques being employed. Here, we focus on the signal processor that takes the phase as the input and generates a sinusoidal amplitude.

* 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.

Nur Fajar Rizqi Annafianto, FNU; Marakkarakath Vadakkepurayil, Jabir; Burenkov, Ivan; Urgurdag, Hasan; Battou, Abdella; Polyakov, Sergey. "FPGA Implementation of a Low Latency and High SFDR Direct Digital Synthesizer for Resource-Efficient Quantum-Enhanced Communication." Presented at IEEE East-West Design and Test Symposium 2020, Varna, BG. September 04, 2020 - September 07, 2020.

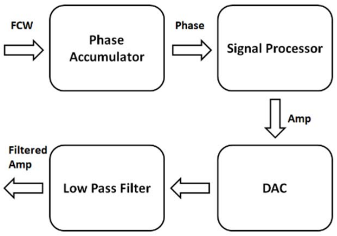


Fig. 1. General block diagram of DDS

CORDIC provides an efficient hardware implementation in terms of area utilization, power consumption and latency. There are three popular approaches to the realization of DDS: (1) Look-Up Tables (LUTs), (2) Polynomial Functions, (namely Tylor series expansion), and (3) a CORDIC algorithm [5].

The LUTs occupy memory, namely, Read-Only Memory (ROM), to store the amplitude of the sinusoidal signal. LUTs are computationally fast, but they require a considerable amount of memory even when compression techniques are used [6]. The memory occupancy is mainly based on the width of FCW input and the width of the output. To get higher spectral purity, the quantization error is minimized by increasing the LUT memory widths of the output amplitude to yield higher precision. The consequence is that memory size grows significantly with the greater spectral purity requirement. In turn, more memory results in higher power consumption, slower operation, and lower stability [7].

The Taylor series expansion has a complicated implementation that uses several multipliers/dividers. To get higher spectral purity, higher-order terms need to be included which in turn increases latency [5].

The CORDIC algorithm calculates sinusoidal amplitude by a set of rotations. The rotational angles in the set are processed in series and the accumulation of the angles approximates the desired angle that corresponds to the necessary output. The number of angles in the set determines the number of iterations, hence the latency. Thus, the correct selection of an angle set is essential. The rotational operation is carried out by adders, logic shifters, and optionally a small amount of memory that makes implementation and integration easier and simpler [3]. For these reasons, CORDIC is better at resource utilization and power consumption.

Spurious Free Dynamic Range (SFDR) defines the spectral purity of the produced signal. The spectral purity of a signal is significant for the overall performance of the system. Since there exists more than one frequency component in a signal, it is necessary to keep the desired frequency's dominance over the spurious components. SFDR is the ratio of the power of the desired signal to the power of the strongest spurious signal. Thus, the higher SFDR the smoother the output obtained, which is preferred and pursued in this work.

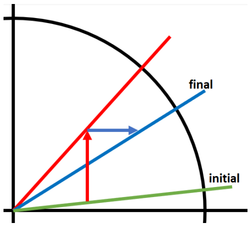


Fig. 2. Rotation mode of CORDIC in DDS with two angular steps

CORDIC has two functional modes: a vector mode and a rotation mode. Our DDS algorithm is based on the rotation mode. In this mode, the initial vector experiences several rotations in cartesian coordinates based on the angle set to reach the desired vector position that corresponds to the destination phase. Figure 2 shows the rotation mode with two phases in the set. In vector mode, the destination angle is estimated by using a set of pre-specified vectors as the reversal of the rotation mode, where it uses the vectors to approximate the target angle [5]. However, CORDIC also comes with some drawbacks. First, it needs scale factor compensation due to numerical operations in the algorithm. Usually, the result is achieved by dividing the scale factor with the output of the series of rotation. To eliminate this requirement, the initial vector is arranged such that it has already been regulated (pre-divided) with the scale factor prior to the calculations. Secondly, accuracy restriction: the number of rotations and the selection of the angles set impacts how close the final angle is to the desired angle [8]. A smaller angle set is beneficial. The following sections describe components in each stage of the hardware architecture and the mathematical operations that they employ.

A. Conventional CORDIC

The first CORDIC algorithm was proposed by removing the burdensome cosine and sine functions multiplications with logic shift operations. Equation (1) computes a rotation of the initial vector (x, y) to the vector (x', y') with the angular distance of θ.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{1}$$

$$\theta = \sum_{i=0}^{b-1} \alpha_i \tag{2}$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \prod_{i=0}^{b-1} cos\alpha_i \begin{bmatrix} 1 & -d_i \tan\alpha_i \\ d_i \tan\alpha_i & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{3}$$

The angle set of α_i converges to θ with a combination of clockwise and/or counterclockwise rotations, see equation (2).

Substituting (2) into (1) and taking $cos\alpha_i$ out of the matrix, we obtain equation (3), where d_i corresponds to the direction of rotation at the respective stage i, given as $d_i = \{-1, 1\} = sign(z_i)$, -1 is for counterclockwise direction and 1 is for clockwise direction. b is the angle set size and the number of iterations. z_i is the remaining phase at iteration i. Our goal is to establish a recurrent formula for rotations that can be conveniently calculated on an FPGA.

$$\alpha_i = \arctan(2^{-i}) \tag{4}$$

$$\prod_{i=0}^{b-1} cos\alpha_i = K \tag{5}$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = K \begin{bmatrix} 1 & -d_i * 2^{-i} \\ d_i * 2^{-i} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \qquad (6)$$

$$z_i = \theta - \textstyle\sum_{c=i}^{b-1} \alpha_c \qquad (7)$$

K in equation (5) is the overall scale factor that can be pre-calculated for the initial vector (x, y). Substituting equation (4) and (5) into equation (3), we obtain equation (6). The division by 2^i can be replaced by an arithmetic shift operator. Thus, iterations are written as:

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} 1 & -d_i * 2^{-i} \\ d_i * 2^{-i} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \qquad (8)$$

The block diagram in Figure 3 shows three stages (i-1), (i), and (i+1) of a conventional CORDIC as the realization of equation (8). The multiplexers have +/- tags that determine the additions or extractions of variables based on the sign of z_i. Note that intersections of lines show the crossing of paths with no connection between them, this is valid for all diagrams.

Equation (4) implies an angle in the angle set for the iteration i. For the sake of simplicity, by selecting 7 iterations ($i = [0, 6]$), we obtain the following angle set: {45, 26.565, 14.036, 7.125, 3.576, 1.789, 0.895}. Note that the conventional CORDIC has 15 iterations. To ensure that z_i is approximately 0 at the end for any destination angle θ, the number of iterations (b in equation (3)) is specified as 15, hence i ranges from 0 to 14. This number is also the accuracy limit of the digital system which uses variables with a bit width of 16 bits, because shifting more than 15 bits results in 0 in such a variable. Thus, these additional variables have no impact on the algorithm, their operations add redundant latency and produce no modification on the final output. However, the range of convergence, that specifies the absolute value of the angle θ, is 99.882. One uses a domain folding technique with 2 blocks that cover [-90, 90) and [90, 270). This way any θ can be reached by locating an appropriate initial vector.

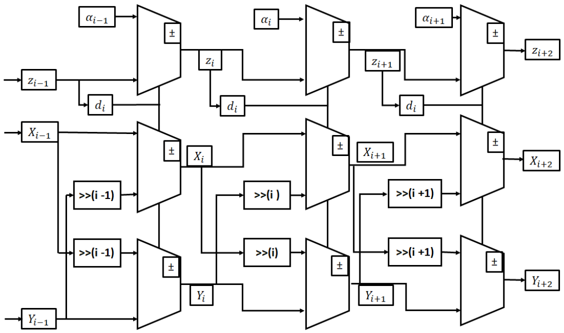


Fig. 3. Block diagram of Conv DDS

B. *Scaling Free CORDIC*

As the name implies, Scaling Free CORDIC pursues an algorithm to avoid multiplication by scale factor prior to the final output for fast performance. Scaling-free CORDIC recognizes one direction of rotation and a halting state, meaning $d_i \in \{0, 1\}$. It rotates counterclockwise only when z_i is greater than the angle at stage i or stays at the current position otherwise. Thus, z_i is always a positive number. This makes the attainable maximum frequency higher as we will see in the result section. The sine and cosine terms can be simplified when any of them is considerably small.

$$\left\lceil \frac{w - log_2 6}{3} \right\rceil \le j \le w - 1 \qquad (9)$$

$$\begin{bmatrix} sin\theta \\ cos\theta \end{bmatrix} = \begin{bmatrix} 2^{-i} \\ 1 - 2^{-(2i+1)} \end{bmatrix} \qquad (10)$$

The approximation in equation (10) is accurate if the requirement in equation (9) is met [4], where w is the bit width. By substituting (10) and (2) into (1) we obtain:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \textstyle\prod_{i=0}^{b-1} d_i \begin{bmatrix} 1 - 2^{-(2i+1)} & -2^{-i} \\ 2^{-i} & 1 - 2^{-(2i+1)} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \qquad (11)$$

Then, the recursive formula is given by:

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = d_i \begin{bmatrix} 1 - 2^{-(2i+1)} & -2^{-i} \\ 2^{-i} & 1 - 2^{-(2i+1)} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \qquad (12)$$

Note that (12) has no scale factor unlike in (6). Figure 4 depicts the block diagram of the Scaling Free CORDIC algorithm at stage i.

The low range of convergence uses a domain folding technique to squeeze the blocks into several extra regions. Due to the condition in (9), and with bit width (w in equation (9)) of 16, where j is i+1, the approximation in (10) only holds for i between 3 to 14, but after the 8th iteration, the logic shifter of (2i+1) results in more than 17 bits shift. Thus, iterations 9 through 14 have no effect. Therefore, iterations 9 through 14 are omitted, to reduce latency and redundant area usage. Hence i goes between 3 to 8. The range of convergence becomes [0, 22.5). The low range of convergence requires extensive use of a domain folding technique. To obtain convergence, 16 domain folding is employed and requires multiplication by a bothersome factor of $1/\sqrt{2}$. Thus, each domain's distance is 20 degrees which is within the range of convergence.

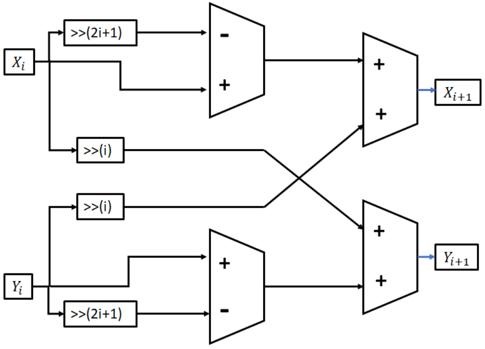


Fig. 4. Block diagram of Scaling-Free CORDIC

This argument reduction technique improves the latency of the method. Particularly, one may jump over several stages by predicting the output at the end of the skipped stages [4]. Typically, more than one computational path is possible, particularly at the early stages. This is because the initial angles are larger than that of the last stages. These computational paths can be pre-computed, and the results can be assigned using multiplexers. Probable combinations of output in the earlier stages are still few and can be estimated by using multiplexers. Hence the first 3 stages are skipped and the output at the end of the 3rd stage is obtained. However, the argument reduction technique requires a variadic scale factor, therefore scale factor multiplication is not avoided completely [4]. Nonetheless, this technique reduces a significant amount of latency (from 12 to 9 iterations as we see in the result section for state-machine based design). The consequence of not reaching the desired angle due to the insufficiency of the range of convergence is to repeat iterations for the rest of the angular gap. The angular gap means the remaining angle to the desired angle that the range of convergence could not cover. Thus, double and even triple latency may occur. Domain folding and the argument reduction techniques are critical in this regard.

The angle set is {36.87/16.26/0, 7.125/0, 1.789, 0.895, S*0.112} where S is an integer in a range from 0 to 8 [9]. The range of convergence is (-57.57, 57.57). Thus, to cover the entire space, quadrant domain folding can be adopted. Domain folding occurs at the first stage. The computation of each phase assumes different strategies.

Friend angles: Any group of angles that have identical magnitude is considered friend angles [9]. For instance, in Cartesian coordinate R = 4 + 3i with the phase of 36.869 and R = 5 with the phase of 0 are friend angle because they have the same magnitude of 5. Thus, all angles in Figure 2 are friend angles since they all have the same magnitude of 1. The identical magnitude is essential for the consistency of the system because different magnitudes impose divergence in power gains and result in different scale factors that make the system even more complicated.

Redundant CORDIC: A Conventional rotator moves a vector in either direction: clockwise or counterclockwise. However, rotation with a large angular gap may require the next angles to cover up the unnecessarily extensive jump in a reverse direction. In those cases, holding the position instead of rotating is advantageous. Thus, the direction of rotation choices are $d_i = \{-1, 0, 1\}$. However, adding one more "direction", that is 0 or no angular movement, yet regulated with appropriate power gain to attain consistency with the other directions, reduces the maximum frequency of the design.

Nanorotator: Rotation by a sufficiently small angle can be approximated further. Given $R = A + Si$, a rotation is sufficiently small if S<<A, therefore $\alpha = \arctan(S/A) \approx S/A$. The other rotators are the same as previously explained for CORDIC algorithms.

III. Implementation

To ensure a successful implementation, we have chosen the workflow depicted in Figure 5. We implemented the algorithm as a MATLAB script and simulated the code, taking advantage of functions that ultimately are not feasible in the hardware platform, such as floating-point, exponentiation, and numerous other operators. This reduces design effort and completion time. Then, we verify the result and evaluate the performance in the software domain, which gives us insight into the possible performance in the hardware domain. We write the register transfer level (RTL) implementation of the design in Xilinx ISE using Verilog HDL. Then, we designed the testbench to simulate the RTL implementation and confirm its functionality. Since all variables in the hardware are integer, we map the hardware simulation results to that of MATLAB simulation for consistency. The hardware platform we utilized is the Xilinx Virtex-6 ML605 FPGA. Next, we verify the FPGA's functionality using an Integrated Logic Analyzer (ILA). We store and extract the output values from the ILA signal analyzer, compare the results with that of RTL simulation, and assess the data for evaluation.

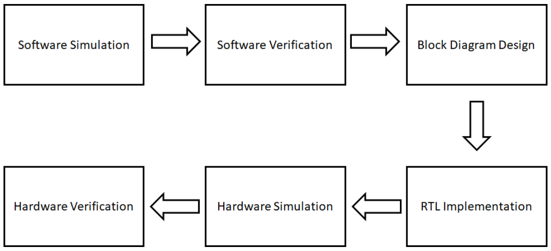


Fig. 5. Workflow

Here, we describe the stages of our implementation:

Stage 1: The domains folding technique, using 8 domains, retains resource efficiency for the system. Additionally, a higher maximum frequency is achieved using one-directional rotations. Folding the coordinate space into several domains leads to a smaller convergence range, but when the number of domains reaches or exceeds 16, complicated operations such as multiplication by $1/\sqrt{2}$ are required. Thus, we use 8 domains to ensure simplicity. The assignment of the initial vector can be done by trivial swapping between imaginary and real parts and negation as seen in Table 1. The angular range of each domain is 45 degrees, which is within the convergence range of the angle set in the counterclockwise direction: it enables one-directional rotation for the next stage.

Table 1. Eight domain folding coordinate assignment

Domain	X	Y
0-45	X	Y
45-90	Y	X
90-135	-Y	X
135-180	-X	Y
180-225	-X	-Y
225-270	-Y	-X
270-315	Y	-X
315-360	X	-Y

Stage 2: The first rotation yields 3 phase options with angles {36.87, 16.26, 0}. All rotation coefficients have the same magnitudes that imply the same power gain/scale factor. We use coefficients, with an angular magnitude of 1.5625. Hence, R = 1.25 + 0.9375i for phase of 36.869 degrees, R = 1.5 + 0.4375i for phase of 16.26 degrees, and R = 1.5625 + 0.0i for phase of 0 degrees. Equation (12) is modified to optimize the hardware implementation for the above three angles such as:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 + 2^{-2} & -1 + 2^{-4} \\ 1 - 2^{-4} & 1 + 2^{-2} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2^{-1} + 1 & -2^{-1} + 2^{-4} \\ 2^{-1} - 2^{-4} & 2^{-1} + 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad (14)$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 2^{-1} + 1 + 2^{-4} \\ 2^{-1} + 1 + 2^{-4} \end{bmatrix}^T \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \quad (15)$$

Equations (13), (14), and (15) can be implemented with just logic shifter and adder.

Resource sharing eliminates redundancy in resource usage. In Figure 6, we use 6 logic shifters as some operators share the same logic shifter's output. The switching rules for the multiplexers are shown as numbers {0, 1, 2}, where {0, 1, 2} encodes the jump angles {36.87, 16.26, 0}, respectively. The architecture of stage 2 is somewhat complex, which may impact the maximum frequency of the hardware implementation. Thus, having a one-directional rotation approach shortens the longest path of the architecture, and in our case, we have 3 rotational options instead of 5 in the regular mode, which shrinks the area usage and improves the speed of this segment.

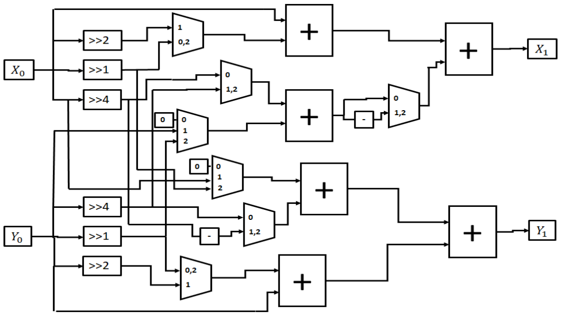


Fig. 6. Stage 2 block diagram

Stage 3: In this stage, we adopt redundant CORDIC to eliminate several rotations and guarantee convergence provided by the remaining angles in the set. The coefficients of this rotator have an angular magnitude of 1.0078125: R = 1 + 0.125i for a phase of 7.125 degrees, and R = 1.0078125 + 0.0i for a phase of 0 degrees. Equation (12) turns into:

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & -d * 2^{-3} \\ d * 2^{-3} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad (16)$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 + 2^{-7} \\ 1 + 2^{-7} \end{bmatrix}^T \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad (17)$$

Hardware implementation of equations (16) and (17) requires just 2 logic shifters per coordinate. The direction d in (16) can be {-1, 1}. The rotation by 0 degrees (described by equation (17)) is equivalent to a no-rotation choice. Such redundancy is tolerable because we end up with three jumping options similar to that of the previous stage. No degradation in the maximum frequency of the design results from this architecture in that regard. The coefficients in stages 2 and 3 ensure consistency of the scale factor as the friend angle's condition is fulfilled.

The block diagram for this stage, Figure 7, shows 4 multiplexers where two of them have the tag numbers. Tag 0 indicates the halting condition for no rotation of the current vector. Note that the appropriate power gain is imposed. Tag 1 indicates either clockwise or counterclockwise rotation set by the sign of the active phase z.

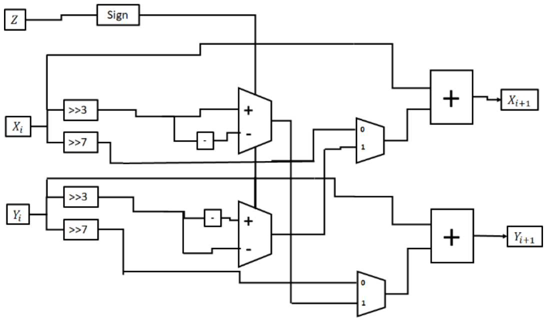


Fig. 7. Stage 3 block diagram

Stage 4: Entering this stage, the residual angle gap's range is 3.58, which is within the range of convergence of the remaining angle in the set. Hence, this stage requires no redundant CORDIC rotation: $d = \{-1, 1\}$. In this stage, we adopt conventional CORDIC architecture at the 5th iteration. The coefficient is R = 1 + 0.03125i for a phase of 1.789 degrees, and the hardware compatible computation is given by equation (18):

$$\begin{bmatrix} x_{i+1} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} 1 & -d * 2^{-j} \\ d * 2^{-j} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (18)$$

The hardware implementation requires two shifters, see Figure 8. For this stage i = 3 and j = 5.

Stage 5: We reuse conventional CORDIC architecture similar to the previous stage but with R = 1 + 0.015625 for a phase of 0.895 degrees. The hardware compatible computation is also given by equation (18), but here i = 4 and j = 6. The block diagram is identical to that of stage 4, which is shown Figure 8.

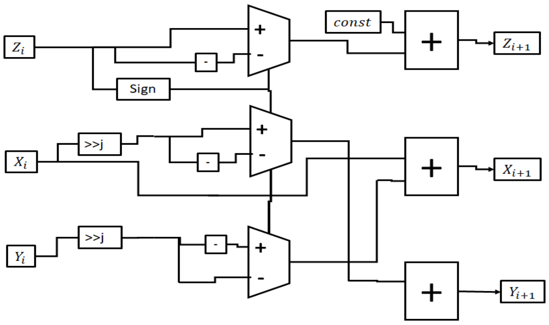


Fig. 8. Stage 4 and 5 block diagrams

Stage 6 (last stage): We are left with a residual angle gap, whose range is 0.875. For this reason, the rotator takes advantage of a nanorotator approximation with a non-constant, adaptive scaling coefficient: R = 1+ (S * 0.001953125i) for variadic phase, where S ∈ [0,8]. Considering the allowed values of S, the range of convergence is (-0.895, 0.895). The hardware compatible version of the coefficient is given in equation (19) and its architecture is shown in Figure 9. The stage 6 implementation requires extra logic: a scale decoder and an attenuation block.

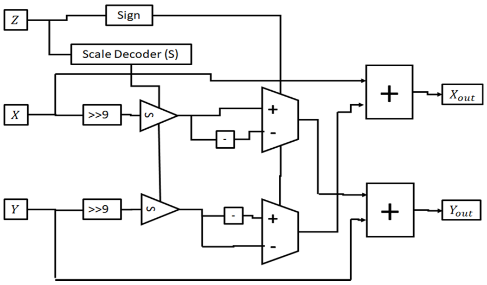


Fig. 9. Stage 6's block diagram

$$\begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & -d * 2^{-9} * S \\ d * 2^{-9} * S & 1 \end{bmatrix} \begin{bmatrix} x_4 \\ y_4 \end{bmatrix} \qquad (19)$$

The scale decoder block determines the magnitude of the adaptive coefficient of S using the remaining phase, to make the residual of Z as close to 0 as possible. 9 combinations of S are obtained, see Table 2.

Table 2. Remaining angle range for S

Range	S	Range	S
(0, 0.0988]	0	(0. 4998, 0.5987]	5
(0. 0988, 0.1977]	1	(0. 5987, 0.6976]	6
(0. 1977, 0.2966]	2	(0. 6976, 0.7965]	7
(0. 2966, 0.3955]	3	(0.7965, 0.895]	8
(0. 3955, 0.4998]	4		

The attenuation block in Figure 9, depicted as a triangular block with an "S" tag, multiplies the adaptive scale S to the shifted coordinate variables X and Y as defined in equation 19. Here, we use a regular multiplier.

The FPGA implementation integrates all these stages in series to complete the design. Given the combination of coefficients of all stages, the cumulative scale factor is K = 1.5757. Hence, we modify the initial vector in stage 1 by pre-dividing the values by this scale factor, which eliminates the other extra multipliers/dividers.

IV. RESULTS

We evaluate the design's performance by measuring the latency, resource usage, logic operator utilization, SFDR, and maximum frequency. These parameters provide trade-off considerations for a target application with specific requirements. For the sake of comparison, we provide values for three different CORDIC-based DDS implementations with and without a pipeline in the architecture. These are conventional CORDIC, modified Scaling Free CORDIC [3], and our proposed design.

Table 3 shows resource utilization based on the number of LUTs, FFs, and memory for a given target device. Here, the LUTs are Xilinx logic blocks. We use ROM as memory: here its usage is measured in bits. In the table, CORDIC represents the conventional CORDIC (it stores 15 phases for the angle set and assumes 16 bits of variable width). SF-CORDIC stands for the modified Scaling Free CORDIC algorithm. The "P" next to the algorithm's name indicates a pipelined version. The initiation interval of every pipelined algorithm is 1, meaning the module can take inputs every clock cycle with no extra delay.

Table 3. Resource utilization

Algorithm	LUT	FF	ROM
CORDIC	764	84	240
CORDIC P	2506	840	
SF-CORDIC	479	98	96
SF-CORDIC P	835	340	
Proposed	400	140	
Proposed P	498	206	

Table 4 presents the utilization of logic operators. Mult, Add, Comp, Mux, and Shift stand for the multiplier, adder, comparator, multiplexer, and logic shifter. All these logic operators run with variables of 16 bits. The logic shifter accepts the variadic length of shift argument.

Table 4. Logic operator usage

Algorithm	Mult	Add	Register	Comp	Mux	Shift
CORDIC	1	7	5	4	21	2
CORDIC P	1	100	65	19	60	
SF-CORDIC		10	6	14	58	4
SF-CORDIC P		28	63	19	49	
Proposed	2	16	31	20	42	
Proposed P	2	24	68	22	32	

In Table 5, the values of SFDR are specified in dB. We compute the SFDR by using the results obtained from the ILA signal analyzer. Iteration in Table 5 indicates the number of rotations, this number is equal to the number of phases in the set. Latency is specified in the number of clock cycles. It represents the overall delay, in clock cycles, due to iterations and additional strategies such as domain folding and argument reduction techniques.

Table 6 lists the maximum frequency in MHz if implemented on a Xilinx Virtex-6 FPGA. The first column shows the maximum frequency for State-Machine (SM) based DDS and the second one lists that of the pipelined version.

Table 5. SFDR, iteration and overall latency

Algorithm	SFDR	iteration	Latency
CORDIC	92.7394	15	17
SF-CORDIC	56.8218	6	9
Proposed	72.2068	5	6

Table 6. Maximum Frequency

Algorithm	Max frequency	Max Frequency P
CORDIC	180	240
SF-CORDIC	226	354
Proposed	211	251

In terms of resource utilization, the proposed design provides an improvement over the existing designs both in pipelined and SM versions. Conventional CORDIC occupies the largest memory usage due to more angles in the set. Although it uses no memory for the pipelined version, the high number of iterations makes the the increment of the other resources enormous. Memory is no longer needed because

Nur Fajar Rizqi Annafianto, FNU; Marakkarakath Vadakkepurayil, Jabir; Burenkov, Ivan; Urgurdag, Hasan; Battou, Abdella; Polyakov, Sergey. "FPGA Implementation of a Low Latency and High SFDR Direct Digital Synthesizer for Resource-Efficient Quantum-Enhanced Communication." Presented at IEEE East-West Design and Test Symposium 2020, Varna, BG. September 04, 2020 - September 07, 2020.

every stage is implemented separately and is active simultaneously, hence each stage is pre-assigned with a constant angle. SM based SF-CORDIC has the lowest resources compared to our design, but its pipeline-base is less efficient as it employs more iterations than our design. Our SM based design doesn't need any iteration because each stage employs a different type of rotator, which makes the resource usage close to that of a pipeline-based design. Our design occupies the smallest area and provides an energy consumption advantage.

SM based conventional CORDIC has the least overall logic operators, while our proposed design compares positively to the pipelined SF-CORDIC. The pipelined version of a conventional CORDIC uses significantly more logic operators compared to the SM based one because 15 iterations that run on a set of resources are expanded into 15 identical sets of resources. The same explanation holds for the logic operator usage in the SM based and pipelined version of the proposed design. Here, the synthesis tool optimizes the resource allocation by substituting a shifter for a concatenation operator due to constant bit shifting. Consequently, the number of logic shifters may not be the same as shown in the block diagrams.

Low latency is desired to enhance the throughput and efficiency of the communication system because delay in the system slows down quantum feedback - a bottleneck and great challenge for quantum-enhanced communication systems. With a latency of just 6 clock cycles, the proposed design is superior to the other algorithms. Although the number of iterations of SF-CORDIC is very close to that of the proposed design, there is an extra delay of 3 clock cycles due to a required additional compensation to the rotation.

The SF-CORDIC has the highest maximum frequency due to one-directional rotation. Our design has a moderate maximum frequency for its implementation. The conventional CORDIC achieves the highest SFDR value due to the high number iteration.

Our design achieves moderate SFDR yet the lowest latency, with approximately 20 dB SFDR and 64% latency reductions compared to that of the conventional CORDIC design.

V. CONCLUSION

In conclusion, we report a new memory free low latency DDS architecture. Although, complex value computations could be employed to calculate the trigonometric equations, but those calculations are computationally difficult in hardware and energy inefficient. To avoid calculation-related inefficiencies, the common approach is to use a Look-Up Table (LUT) with phase being the input and amplitude being the output. To generate a desired smooth radio-frequency signal small-step quantization is needed, requiring a larger LUT. The LUT requirement leads to an increase in memory usage and may lead to the reduction of the maximum frequency of the FPGA design which would also limit modulation capabilities.

On the other hand, a CORDIC technique offers low complexity and memory-free trigonometric calculation approach, at the expense of extra latencies to complete the computation. We use a pipelined approach to shorten latency. In this design, the sinusoidal wave amplitude is obtained every cycle, in order to maximize our modulation capabilities. To make a quantum measurement enhanced transceiver, we choose the modulation scheme which includes choosing the number of states M, the frequency, and the initial phase detuning between the adjacent states and other communication parameters. All M states are being prepared in parallel at all times, and the active output state is picked according to the encoding and measurement protocols. Because M could be quite large (up to 16 in our implementation) the low-resource usage DDSs are essential for this purpose. In communication links, sensitivity is often measured as the probability to receive an erroneous symbol with certain energy at the receiver. Classical receivers have a sensitivity limit known as the standard quantum limit (SQL). This limit arises from the inevitable shot noise on the idealized classical receiver scheme - a homodyne measurement followed by a perfect detector with no noise of its own and with the 100% detection-efficiency. The SQL is accessible only through quantum measurement. With the help of the described DDS, we have implemented a quantum-measurement telecommunication testbed and demonstrated that the sensitivity of a telecommunication channel is better than the SQL for many different modulation protocols, including quantum-measurement specific modulation protocols, described elsewhere [10].

We intend to use modulation schemes that require a simultaneous phase and frequency modulation. Our novel design achieves the shortest latencies, maximizes modulation capabilities, and uses the minimal footprint compared to other CORDIC-based DDSs.

ACKNOWLEDGMENT

This work is partially supported by National Science Foundation USA through ECCS 1927674.

REFERENCES

[1] I.A. Burenkov, M.V. Jabir, N.F.R. Annafianto, A. Battou, and S.V. Polyakov, "Experimental Demonstration of Time Resolving Quantum Receiver for Bandwidth and Power Efficient Communications", Proc. of *CLEO Conf. on Laser Science to Photonic Applications*, California, USA, 2020.

[2] P. Saravanan and S. Ramasamy, "Sine/cos generator for direct digital frequency synthesizer using pipelined CORDIC processor," Proc. of *Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1-6, Tiruchengode, 2013.

[3] R. Xin, X. Zhang, H. Li, Q. Wang, and Z. Li, "An Area Optimized Direct Digital Frequency Synthesizer Based on Improved Hybrid CORDIC Algorithm," Proc. of *Int. Workshop on Signal Design and Its Applications in Communications*, pp. 243-246, Chengdu, China, 2007.

[4] Y. Xue and Z. Ma, "Design and Implementation of an Efficient Modified CORDIC Algorithm," Proc. of *IEEE Int. Conf. on Signal and Image Processing (ICSIP)*, pp. 480-484, Wuxi, China, 2019.

[5] M.M. Anas, R.S. Padiyar, and A.S. Boban, "Implementation of Cordic Algorithm and Design of High Speed Cordic Algorithm," Proc. of *Int. Conf. on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 1278-1281, Chennai, India, 2017.

[6] Y.S. Gener, S. Gören, and H.F. Ugurdag, "Lossless Look-Up Table Compression for Hardware Implementation of Transcendental Functions," Proc. of *IFIP/IEEE Int. Conf. on Very Large Scale Integration (VLSI-SoC)*, pp. 52-57, Cuzco, Peru, 2019.

[7] W. Shuqin, H. Yiding, Z. Kaihong, and Y. Zongguang, "A 200MHz Low-Power Direct Digital Frequency Synthesizer Based on Mixed Structure of Angle Rotation," Proc. of *IEEE Int. Conf. on ASIC*, pp. 1177-1179, Changsha, China, 2009.

[8] K. Maharatna, S. Banerjee, E. Grass, M. Krstic, and A. Troya, "Modified Virtually Scaling-Free Adaptive CORDIC Rotator Algorithm and Architecture," in *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, pp. 1463-1474, 2005.

[9] M. Garrido, P. Källström, M. Kumm, and O. Gustafsson, "CORDIC II: A New Improved CORDIC Algorithm," in *IEEE Tran. on Circuits and Systems II: Express Briefs*, vol. 63, pp. 186-190, 2016.

[10] I.A. Burenkov, O.V. Tikhonova, and S.V. Polyakov, "Quantum Receiver for Large Alphabet Communication," *Optica*, vol. 5, pp. 227-232, 201

NIST SP 1283
August 2022

Proceedings of the ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2020
August 16–August 19, 2020, St. Louis, MO

DETC2020-19995

ENABLING TRACEABILITY IN AGRI-FOOD SUPPLY CHAINS USING AN ONTOLOGICAL APPROACH

Farhad Ameri
Associate Professor
Engineering Informatics Lab
Texas State University
San Marcos, Texas, USA
ameri@txstate.edu
mailto:alolikanath1@gmail.com

Evan Wallace
Research Scientist
Systems Integration Division
National Institute of Standards and Technologies (NIST)
Gaithersburg, Maryland, USA
Evan.Wallace@nist.gov

Reid Yoder
Research Assistant
Engineering Informatics Lab
Texas State University
San Marcos, Texas, USA
Reid.yoder@ymail.com

ABSTRACT

Traceability of food products to their sources is critical for quick responses to a food emergency. US law now requires stakeholders in the agri-food supply chain to support traceability by tracking food materials they acquire and sell. However, having complete and consistent information needed to quickly investigate sources and identify affected material has proven difficult, and in some cases, costly. There are multiple reasons that make food traceability a challenging task including diversity of stakeholders and their lexicons, standards, tools and methods; unwillingness to expose information of internal operations; lack of a common understanding of steps in a supply chain; and incompleteness of data. Ontologies can address the traceability challenge by creating a shared understanding of the traceability model across stakeholders in a food supply chain. They can also support semantic mediation, data integration, and data exploration. This paper reports an ongoing effort aimed at developing a formal ontology for supply chain traceability using use cases and data from partners in the bulk grain domain. The developed ontology was validated in VocBench environment through creating RDF triples from real datasets and executing SPARQL queries corresponding to predefined competency questions.

Keywords: ontology, supply chain, traceability, traceable resource unit, critical tracking event, interoperability

INTRODUCTION

Traceability of food and feed is becoming an increasing concern among governments, producers, and consumers. Governments wish to act quickly to identify and take tainted food out of the supply chain in response to a food emergency. Producers wish to minimize their exposure to risk and ensure the quality of the food they sell. Consumers are increasingly interested in where their food comes from, what processes were used to produce it, and what it may contain (such as pesticides or genetically modified elements). Traceability can address all these concerns but is challenging to achieve due to the wide range of diverse, disconnected, participants in a supply chain spanning material source to consumer.

Wallace, Evan K.; Ameri, Farhad; Yoder, Reid. "Enabling Traceability in Agri-Food Supply Chains Using an Ontological Approach." Presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC-CIE 2020). August 17, 2020 - August 19, 2020.

The Institute of Food Technologists (IFT) has proposed an approach to address some of these challenges [1]. The approach focuses on a few kinds of occurrences, which they call Critical Tracking Events (CTEs) that are key parts of the lifecycle of a product or of another participant in that product's lifecycle. For traceability, some key lifecycle parts include when material is created, transformed, transferred, changes ownership or custody, changes location, or is consumed or destroyed. The IFT framework also associates Key Data Elements (KDEs) with each type of CTE. These KDEs describe the type of data that should be collected for each event to support later track and trace. In this way the relevant events can be found and assembled, when needed, to determine history related to a product instance. This history can then be queried to answer questions about the product, such as where it may have been contaminated, who may have purchased material from a particular lot of product, or many other questions, the answers to which, could help optimize or improve production or handling of similar future product.

The CTE/KDE approach to traceability requires that only a minimal set of data be captured, and that it be shared when needed to those addressing a health emergency or a business need of the organization providing it. These are all useful characteristics for supporting the challenges in understanding a product history that involves many parties with potentially multiple means and systems for managing their information, particularly since these parties are also reluctant to share production and operation information lest competitors use it to gain advantage. Other issues hampering an end-to-end view of product history has been incomplete data and the use of different IDs, naming conventions, and data formats across systems and partners.

While adopting standards for types of Critical Tracking Events, the data elements that should be captured for them, and identification schemes for related entities would address many of the current challenges for end-to-end traceability data, developing these standards and having them adopted nearly universally across food and agriculture business is both a political and practical challenge. However, the researchers of the Supply Chain Traceability for Agri-Food project at NIST and their partners at Texas State University posit that ontologies and W3C linked data standards and tools may facilitate much earlier impact from the CTE/KDE framework on traceability in the agri-food sector. This is because these standards were designed to support integrating diverse information, and reason over the results of that integration even when that information is incomplete.

Ontologies and the use of W3C standards such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) and tools employing them would address traceability for agri-food in the following ways:

Standardization/Common Understanding: Ontologies can be created that formally define standard Critical Tracking Event (CTE) types and associated Key Data Elements (KDE). This would be used to ensure a shared understanding of these things across stakeholders in a food supply chain.

Data Integrity: The CTEs and KDEs ontologies can be used to specify the completeness and consistency of data that must be present (interpreting ontologies as Integrity Constraints) for the integrity of the traceability system.

Semantic Integration/Mediation: CTE and KDE ontologies can be used as a global model for traceability data to define data forms for common formats. These ontologies can also act as global models for querying data over heterogeneous systems (using a Global and Local As View or GLAV approach, and lifting and lowering patterns) and for integrating the results.

Reasoning for History Exploration, Discovery and Construction: Traceability ontologies supplemented with additional semantic models could be used to support traceability data exploration and "what if" queries to discover important relationships and fill in missing information during a trace-back and trace forward effort related to a food incident.

This paper reports on ongoing efforts to evaluate these hypotheses using use cases and data from partners in the bulk grain domain.

INDUSTRIAL ONTOLOGIES FOUNDARY (IOF)

The Supply Chain Traceability (SCT) Ontology is being developed in conjunction with the Supply chain Working Group activities within the Industrial Ontology Foundry (IOF) initiative. The IOF is an international community of academia, industry, and research institutes that was formed with the vision of increasing the adoption of ontologies in the manufacturing sector [2]. The technical goals of IOF include [3]:

- Create open, principles-based ontologies from which other domain-dependent or application-specific ontologies can be derived in a modular fashion.
- Ensure that IOF ontologies are non-proprietary and non-implementation-specific, so they can be reused in different industrial subdomains and standard bodies.
- Provide principles and best practices by which quality ontologies will support interoperability
- Institute a governance mechanism to maintain and promulgate the goals and principles.
- Provide an organizational framework and governance processes that ensure conformance to IOF principles and best practices.

IOF is particularly focused on developing domain-specific reference ontologies. These reference ontologies can be further extended to create application ontologies. A Reference Ontology in a specific domain is intended to represent the theories and the general knowledge of the domain, independent of particular applications. Domain-specific Reference Ontologies (DSRO) are reused across multiple applications in the domain. IOF ontologies are aligned with a Top-Level Ontology (TLO). Top-level ontologies (a.k.a. upper or foundational ontologies) are highly abstract, domain-neutral ontologies that establish a common framework for creating reference and application ontologies [4]. TLOs provide a broad view of the world suitable for many different target domains. Some of the notable upper-level ontologies include Basic Formal Ontology (BFO) [4], Domain Ontology for Linguistic and Cognitive Engineering (DOLCE) [5], PSL [6], and Suggested Upper Merged Ontology (SUMO) [7]. IOF uses BFO as the TLO in its architecture. BFO has been used widely in the biological domain for integrating disparate ontologies or data models and developing interoperable ontologies for biological applications [8]. There are several reasons that make the investigation of using BFO as TLO worthwhile for many domains including the supply chain domain. Firstly, BFO has a very large user base and it is widely used in a variety of ontologies including military and intelligence. Secondly, BFO is very small, with only 35 classes, and correspondingly easy to use and easy to learn. Additionally, BFO is very well-documented and there are multiple tutorials, guidelines, and web forums for using BFO in ontological projects.

Currently, there are five active working groups (WGs) in IOF. Four of them are addressing different subdomains of manufacturing, including supply chain, production planning and scheduling, maintenance, and product-service systems. The last working group, namely the top-down WG, serves as the glue by providing a common ontology and ensuring consistency across other working groups.

USE CASE DESCRIPTION

The use case discussed in this paper is derived from a proof-of-concept (POC) effort in a project within the agriculture e-business consortium, AgGateway[1], called, Commodity Automation for Rail & Truck (CART). The goal of this project was to facilitate "grain traceability from combine to grain cart, to truck, to elevator, to food processor". The focus of the CART POC in 2017 was on tracking bulk grain from harvest to on-farm storage or from harvest to delivery at a grain elevator. The research discussed in this paper addresses transfer events that take place on the farm in support of these use cases. A more detailed description of the CART project and its POCs is available in [9].

While the IFT CTE movement types often mentioned are shipping and receiving; materials, such as bulk grain, behave somewhat like liquids, so tracking the movement of these materials between containers becomes important for traceability (since such movements can result in difficult-to-reverse mixing with other materials or leave behind trace amounts of material that could intermingle with later loads placed in the same container). The CART project developed an XML Schema component for a movement event type for this notion that it called Transfer Event. As already alluded, a Transfer Event represents an occurrence where a portion of material (the subject of the event) is moved from a source container to a destination container. The notion of *container* involved in these events is abstract, allowing, for example, a harvesting activity to be represented as a Transfer Event from a portion of a field (a container) to a harvester's grain tank (also a container). Thus, the lifecycle of grain being harvested on a farm can be understood to be a series of transfer events. For the harvest to on-farm storage use case in the CART 2017 POC, the sequence of transfer events involved is 1) a harvesting pass, 2) transfer from harvester to cart (a wagon designed to move material within the farm), 3) transfer from cart to an on-farm storage bin. Figure 1 depicts this in schematic form with boxes representing container roles and arrows to signify events.

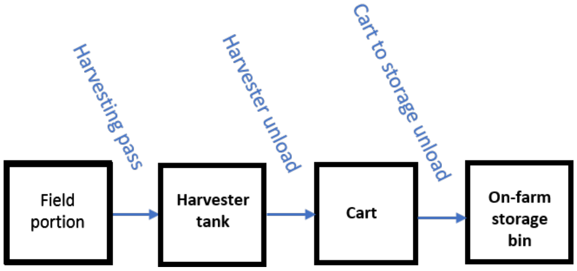


Figure 1. The sequence of Transfer Events involved in the harvest to on-farm storage use case (boxes represent container roles and arrows signify events)

ONTOLOGY DEVELOPMENT METHODOLOGY

Development of SCT ontology follows both top-down and bottom-up approaches. The top-down approach is guided by the IOF architecture that requires IOF ontologies to be aligned to top-level and reference ontologies. Because the traceability ontology is related to the supply chain domain, it is aligned with Supply Chain Reference Ontology (SCRO) [10]. Therefore, both BFO and SCRO were used as imported ontologies at the early stages of ontology development.

A bottom-up approach is also adopted in a sense that a real use case related to the bulk grain domain is selected to be used for requirements definition and ontology validation. As one of the preliminary steps, a set of Competency Questions (CQ) was proposed to validate the ontological content against the use case requirements. This is a common practice in ontology development efforts [11]. The dataset related to the use case together with the information collected from domain experts were used to identify key notions that need to be formalized in the ontology. Once key notions are identified, informal (Subject-Matter Expert) and formal definitions are created for each notion and the necessary relationships and axioms are added to the model. The final step is the creation of an OWL file in Protégé.

Besides following the architecture devised by IOF, the ontology development process in this work conforms to IOF Technical Principles [12] and best practices of ontology development. One of the important rules that the working group has adhered to is

[1] https://www.aggateway.org/

the *True Path Rule* which indicates every instance of a child class must also be an instance of every class that is a parent of the child class. Applying this rule ensures that *multi-inheritance* is avoided in the asserted taxonomy of the ontology.

SUPPLY CHAIN TRACEABILITY (SCT) ONTOLOGY

The Supply Chain Traceability (SCT) ontology is intended to serve as the canonical model for traceability data in agri-food supply chains. The SCT ontology can be used for formal representation of Critical Tracking Events and their associated Key Data Elements (KDE), as well as the entities that participate in those events, including the subjects of traceability efforts that are referred to as Tracking Resource Units (TRU). The SCT ontology should also provide the means for timestamping the tracking events and linking them to the geospatial regions they occur at. Using the SCT ontology, the traceability graph of the supply chain can be represented formally (Figure 2) and traversed wholly or partially in order to reconstruct the history of the material entities that flow through different branches of the graph in different temporal intervals.

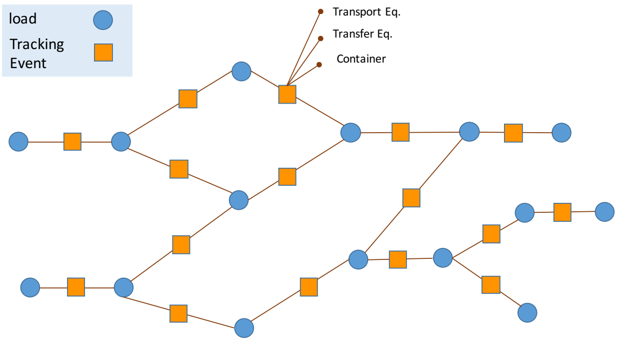


Figure 2. Traceability graph: by traversing the graph, the history of the traceable units can be reconstructed.

Three main modules of the SCT Ontology, namely, Traceable Resource Unit, Critical Tracking Event, and Container are described the following sections. A brief introduction of BFO types is provided first.

BFO Classes

BFO (Basic Formal Ontology) splits all entities into two categories: *continuants* and *occurrents*. Continuants are the entities that continue to persist through time while maintaining their identity. BFO recognizes a dichotomy between *independent* and *dependent* entities. Grains and containers, for example, are independent continuants whereas a *quality* of a container (for example, its mass) is a dependent continuant since this mass is dependent on the container. If the container ceases to exist, then so also would its mass. Occurrents are the processes, events, or happenings in which continuant entities participate. In the traceability use case, all tracking events are considered to be *occurrents*. Another BFO class that is used extensively in SCT ontology is the *role* class. Role is an entity that is realized (manifested or actualized) in a process. Examples include the role of an organization to serve as a supplier in a supply chain or the role of a person as a truck driver. A supplier's roles are realized in the supply chain processes in which the organization participates. Roles are dependent continuants as they can exist only insofar as they are roles of some independent continuants. In the class diagrams in the following sections, green boxes denote BFO classes and blue boxes represent SCRO (Supply Chain Reference Ontology) or SCT classes.

Traceable Resource Unit

Traceable Resource Units are collections of material with some shared history for which some agent may have a need to retrieve information. The shared history might include similar production, movements, or storage history. Some subtypes of TRU include lot, load, batch, and shipment. In SCT Ontology, TRU is treated as a defined class rather than as an asserted universal. A *universal* is an entity that can exist on its own rights and is part of the official (asserted) *is-a* hierarchy. Examples include a *portion of corn* that is considered to be a type of BFO:material entity or the act of harvesting that is a BFO: occurrent. A *defined class*, on the other hand, is composed from classes, individuals, and relations using equivalence axioms. Any instance of object or object aggregate that bears a TRU Role is classified as an instance of TRU by the reasoner. One important type of TRU related to the motivating use case in this work is Load. A Load is a collection of material transferred or transported together. The semi-formal definitions of Load and Load Role in SCT Ontology are provided in Table 1.

Table 1: Definitions for Load and its related roles

Notion	Definition
Load	A BFO: object or BFO: object aggregate that bears a load role
Load Role	A BFO: role that inheres in an object or object aggregate when they are transferred or transported together.
Source Load Role	A role that inheres in objects or object aggregates when they participate as the input material in a transfer event.
Target Load Role	A role that inheres in objects or object aggregates when they participate as the output material in a transfer event.

As shown in Figure 3, a portion of corn, that is asserted to be an instance of material entity, can be inferred to be an instance of the Load class as well since it is the bearer of a Load Role that is realized in some Transfer Event. Since Load

Role is a realizable entity, it ends when the process that realizes the role ends. Two instances of load (L1 and L2) can be combined, through some transfer event (TE1), into a third instance of load (L3). TE1 is the entity that links L1 and L2 to L3. The loads that are input to a transfer event are inferred to be instances of Source Load and the new load created through the event is an inferred instance of Target Load as shown in Figure 7.

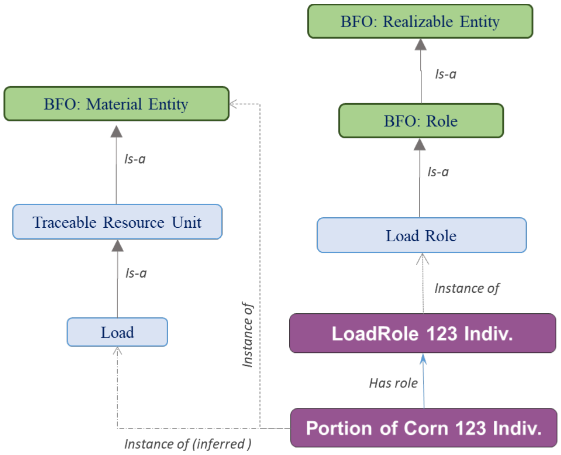


Figure 3. Using inference for identifying Load individuals

Critical Tracking Event

The Critical Tracking Events (CTEs) are the actual events, or processes (BFO: occurrent), that occur to the traceable units during their lifecycle, such as receiving, transferring, transforming, packing, shipping, and transporting. In SCT Ontology the CTE class is defined as an occurrent which has at least one TRU as a participant. To enable end-to-end supply chain traceability, ideally, all CTEs should be identified and recorded since those events are the key elements that contain the history of the supply chain. CTE records are used to reconstruct the history of TRUs later during trace-back or trace-forward analysis. Different types of Tracking Events in SCT Ontology include Transfer Event, Transport Event, Transformation Event (including Drying or Blending), and Ownership Change Event.

Transfer Event is a type of Movement Event that involves moving the subject material from a source container to a target container. For example, moving 100 pounds of soybeans, using a conveyor belt, from one bin to another bin, is an example of a transfer event. A transfer event typically has different participants such as operator, transfer device, and container as shown in Figure 4.

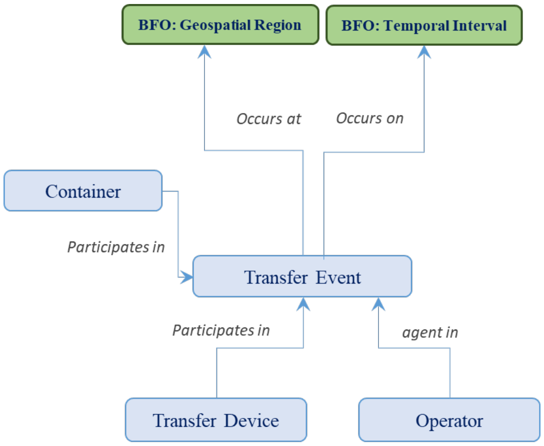


Figure 4. Different types of participants in a transfer event

Timestamping of Tracking Events in SCT Ontology is conducted using the pattern shown in Figure 5. A Transfer Event *occurs on* a Temporal Interval (one-dimensional- temporal region) which is a type of BFO: Temporal Region. A Temporal Interval has a beginning and an ending instant (BFO: Temporal Instant) which are designated by Time of Day Identifiers with xsd:dateTime values. Time of Day Identifier is a Designative Information Content Entity which is an imported class from Information Artifact Ontology (IAO) [13].

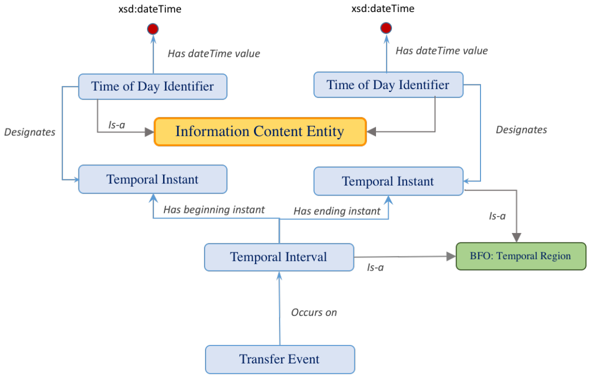


Figure 5. Timestamping of Transfer Event

The IAO is a domain-neutral ontology for representing information entities that stand in a relation of aboutness to continuants and occurrents. The location of Tracking Events is captured using *occurs at* property that has Geospatial Region in its range. The Geospatial Region is designated by a Global Location Number (GLN) that is type of IAO: Geospatial Region Identifier.

The location can also be specified indirectly by specifying the `Facility` in which the event occurs.

Container

In SCT Ontology, `Container` is regarded as a role class (defined class). Any instance of material entity that bears a `Container Role` can be classified by the reasoner as a container individual. For example, the `Container Role` can be inhered in the `Field` itself. A `Container Artifact`, on the other hand, is actually an artifact that is designed and intended to contain material. In this use case, the container individuals are instances of `Container` class (the defined class) to provide more flexibility in treating different entities as containers.

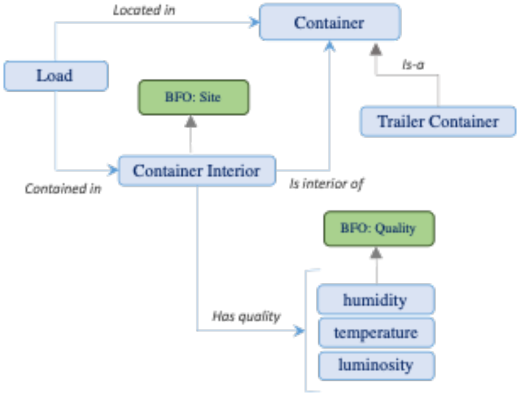


Figure 6. Container and its relationship with Load and Container Interior

The sub-classes of `Container` include `Combine Tank Container`, `Grain Bin Container`, `Trailer Container`, and `Railcar Container`. A Container can have several qualities such as weight, height, and volume. In some occasions, it is beneficial to separate the interior of the container from the container itself because the interior might have different qualities (such as humidity and temperature) that need to be recorded. For this reason, `Container Interior` (a sub-class of `BFO: site`) is included in the ontology as shown in

Figure **6**.

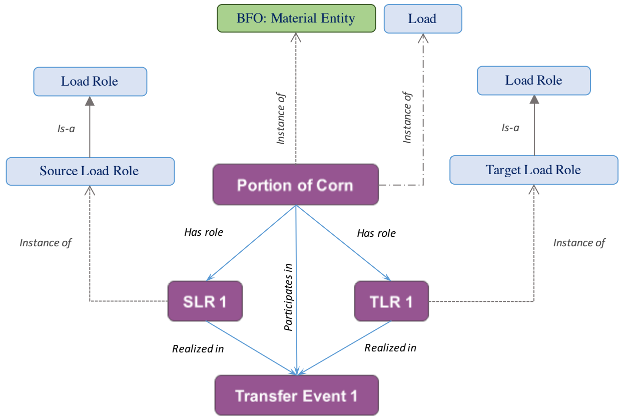


Figure 7. Instantiation of Source and Target Loads

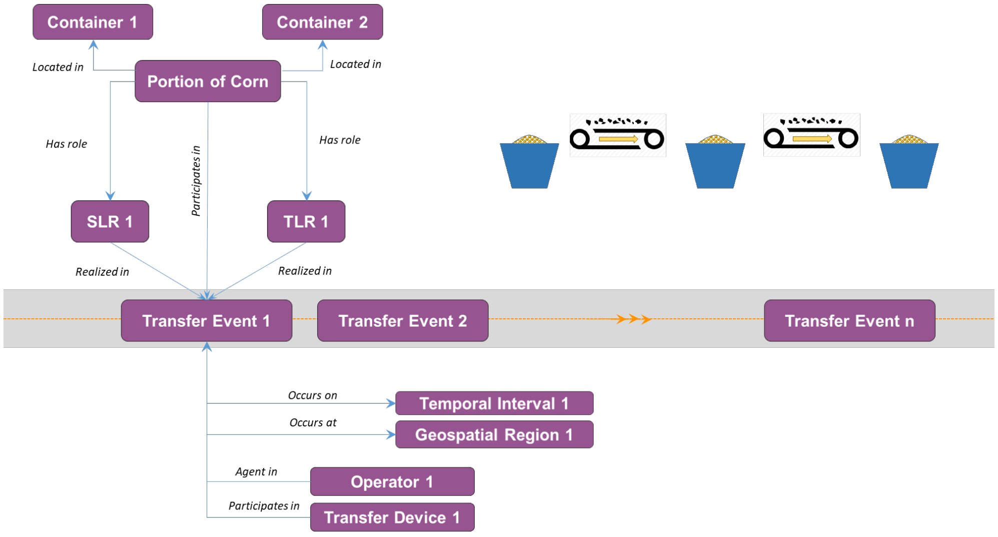


Figure 8. The instance model related to the baseline use case with Transfer Events Only

Wallace, Evan K.; Ameri, Farhad; Yoder, Reid. "Enabling Traceability in Agri-Food Supply Chains Using an Ontological Approach." Presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC-CIE 2020). August 17, 2020 - August 19, 2020.

IMPLEMENTATION AND VALIDATION

VocBench [14] was used as the environment for creating RDF triples (knowledge graph) from the raw data, provided by external sources, and executing SPARQL queries. A real dataset in .xlsx format was used for validation purposes. It should be noted that the data scheme of the .xlsx file did not conform to the ontology structure. In fact, it is the role of the ontology to serve as the unifying framework for multiple heterogenous datasets and harmonize and integrate them as a uniform RDF dataset.

Validation was conducted is two steps. In the first step, a simple scenario with only 3 sequential Transfer Events, as shown in Figure 8, was explored to create a baseline. The instances for this scenario were directly created in Protégé and then transferred into VocBench environment. Since the answers to the competency questions for this baseline scenario are known, it can be readily verified if the ontology is logically consistent and the CQs can be properly answered. The CQs used for the baseline scenario are listed below:

1. What are the containers used in transfer event 2?
2. What are the transfer events related to this load (portion of corn)?
3. What are the containers that this load has been in contact with?
4. What was the location of this load on a certain date/time?

The second step of validation involved running more sophisticated queries against the real-life dataset. In the following section the triplication and query formulation processes are described in further details.

Triplication in VocBench

VocBench is a free and open-source RDF modelling platform realized through a collaborative and multilingual web-based environment. With support for OWL ontologies, SKOS thesauri, and other RDF datasets, VocBench allows users to edit, manage, and transform datasets through various embedded tools and processes. Two valuable tools VocBench offers are Sheet2RDF—a comprehensive interface for acquiring and transforming RDF triples from external datasheets (.xlsx, .csv, etc.)—and a SPARQL query engine capable of querying both explicit and inferred RDF triples, allowing users to answer high-level questions about different datasets.

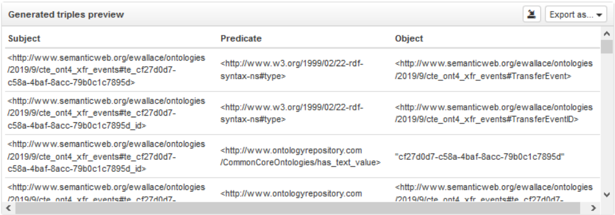
The driving force behind Sheet2RDF is PEARL, a triplication language that parses the uploaded datasheet and maps the information to the dataset. For our purposes, we've utilized PEARL's capabilities to gather data from transfer event spreadsheets and map the information to the traceability ontology. Figure 9 shows the PEARL code was executed iteratively for every row in a *transfer event datasheet* to create new transfer event instances in the ontology. The code is divided into two sections: *nodes* and *graph*. In the nodes section, RDF nodes are created from the data contained in the current datasheet row. More specifically, relevant RDF literals are created from the data, along with the generation of any necessary URIs. In the graph section, the RDF nodes created in the nodes section are used to define the RDF triples in relation to the ontology. More specifically, an individual of the `Transfer Event` class is created, along with the relevant properties. Once the PEARL code is executed, the generated RDF triples are partially shown in Figure 10. From here, the user has the option of importing the triples into the ontology, or exporting the triples externally.

```
nodes = {
  subject
uri(<http://infoneer.txstate.edu/ConcatConverter>("
te_", "")) col_1/value .
  te_id_val literal^^xsd:string col_1/value .
  te_id
uri(<http://infoneer.txstate.edu/ConcatConverter>($
subject, "_id")) .
  te_desc_val literal^^xsd:string col_5/value .
  te_desc
uri(<http://infoneer.txstate.edu/ConcatConverter>($
subject, "_description")) .
  te_inter
uri(<http://infoneer.txstate.edu/ConcatConverter>($
subject, "_interval")) .
  te_start_time_val  literal^^xsd:string  col_8/value
.
  te_start_time
uri(<http://infoneer.txstate.edu/ConcatConverter>($
te_inter, "_start_time")) .
  te_start
uri(<http://infoneer.txstate.edu/ConcatConverter>($
te_inter, "_start")) .
  te_end_time_val literal^^xsd:string col_12/value .
  te_end_time
uri(<http://infoneer.txstate.edu/ConcatConverter>($
te_inter, "_end_time")) .
  te_end
uri(<http://infoneer.txstate.edu/ConcatConverter>($
te_inter, "_end")) .
}
graph = {
  $subject rdf:type :TransferEvent .
  $te_id rdf:type :TransferEventID .
  $te_id core:has_text_value $te_id_val .
  $subject core:designated_by $te_id .
  $te_start_time rdf:type core:TimeOfDayIdentifier .
  $te_start_time                 core:has_text_value
$te_start_time_val .
  $te_start rdf:type bfo:BFO_0000148 .
  $te_start core:designated_by $te_start_time .
  $te_end_time rdf:type core:TimeOfDayIdentifier .
  $te_end_time  core:has_text_value  $te_end_time_val
.
  $te_end rdf:type bfo:BFO_0000148 .
  $te_end core:designated_by $te_end_time .
  $te_inter rdf:type core:TemporalIntervalIdentifier
.
```

Wallace, Evan K.; Ameri, Farhad; Yoder, Reid. "Enabling Traceability in Agri-Food Supply Chains Using an Ontological Approach." Presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC-CIE 2020). August 17, 2020 - August 19, 2020.

```
  $te_inter :hasBeginningInstant $te_start .
  $te_inter :hasEndingInstant $te_end .
  $subject :occursOn $te_inter .

  OPTIONAL {
    $te_desc rdf:type :TransferEventDescription .
    $te_desc core:has_text_value $te_desc_val .
    $subject core:described_by $te_desc .
  }
}
```

Figure 9. VocBench screen showing the PEARL code for creating new Transfer Events instances in the ontology from the datasheet.



Generated triples preview

Subject	Predicate	Object
<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#te_cf27d0d7-c58a-4baf-8acc-79b0c1c7895d>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#TransferEvent>
<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#te_cf27d0d7-c58a-4baf-8acc-79b0c1c7895d_id>	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#TransferEventID>
<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#te_cf27d0d7-c58a-4baf-8acc-79b0c1c7895d_id>	<http://www.ontologyrepository.com/CommonCoreOntologies/has_text_value>	"cf27d0d7-c58a-4baf-8acc-79b0c1c7895d"
<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#te_cf27d0d7-	<http://www.ontologyrepository.com	<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#te_cf27d0d7-

Figure 10. The partial view of the generated RDF Triples.

Formulating and Executing SPARQL Queries

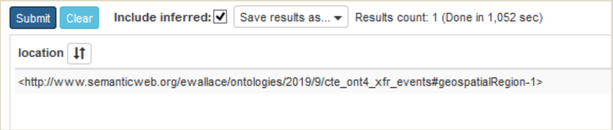
Once the user has incorporated any necessary RDF triples into the ontology, SPARQL queries can be formulated and executed in order to understand specific characteristics about the dataset. In the following example query shown in Figure 11, the question "What was the likely location of a given load at a given datetime?" is answered. Specified through the BIND statements at the beginning of the query, the user declares the individual "`:portion-of-corn-123`" from the ontology as the given load, and "`2019-10-26T22:20:00`" as the datetime that must occur during any events of interest. The rest of the statements in the query serve to traverse the structure of the ontology to find the location solution through the following steps:

1. Find the transfer events the specified load was involved in
2. Acquire the beginning and ending datetimes of said transfer events
3. Check which transfer events the specified datetime lies within
4. Return the location(s) said transfer events occur at

```
# What was the likely location of a given load at a given
datetime?

SELECT DISTINCT ?location WHERE {
    BIND(:portion-of-corn-123 AS ?load) .
    BIND("2019-10-26T22:20:00"^^xsd:dateTime AS
?date_time) .

    # ensure specified load was involved in transfer event
    ?load bfo:RO_0000087 ?load_role .
    ?load_role rdf:type/rdfs:subClassOf* :LoadRole .
    ?load_role bfo:BFO_0000054 ?transfer_event .
    ?transfer_event rdf:type/rdfs:subClassOf*
:TransferEvent .

  # get transfer event beginning and ending datetimes
    ?transfer_event core:occurs_on ?interval .
    ?interval rdf:type/rdfs:subClassOf* bfo:BFO_0000038 .
    ?interval :hasBeginningInstant ?beginning .
    ?beginning rdf:type/rdfs:subClassOf* bfo:BFO_0000148 .
    ?beginning core:designated_by ?beginning_id .
     ?beginning_id rdf:type/rdfs:subClassOf*
core:TimeOfDayIdentifier .
    ?beginning_id core:has_datetime_value
?beginning_literal .
    ?interval :hasEndingInstant ?ending .
    ?ending rdf:type/rdfs:subClassOf* bfo:BFO_0000148 .
    ?ending core:designated_by ?ending_id .
    ?ending_id rdf:type/rdfs:subClassOf*
core:TimeOfDayIdentifier .
    ?ending_id core:has_datetime_value ?ending_literal .

# ensure specified datetime is within transfer event
beginning and ending datetimes
    FILTER(?date_time >= ?beginning_literal) .
    FILTER(?date_time <= ?ending_literal) .

  # get transfer event location
    ?transfer_event :OccursAt ?location .
     ?location rdf:type/rdfs:subClassOf*
     core:GeospatialRegion .
} LIMIT 100
```

Figure 11. The SPARQL query related to finding the likely location of a given load at a given time.

The query is executed, and the results are shown in Figure 12. In this case, there is one location retuned as the response. Thus, it is realized that "`:portion-of-corn-123`" was in "`:geospatialRegion-1`" in a time interval that included the specified datetime "`2019-10-26T22:20:00`". Note that in this example, "`:portion-of-corn-123`" has two types: 1) `Portion of Grain` (asserted type) and 2) `Load` (inferred type). The `Load` class is not directly instantiated in this case. However, in large knowledge graphs for enterprise systems where reasoners are not capable of handling a huge number of inference tasks efficiently, the practice of directly instantiating classes (such as `Load`) with equivalence axioms is common.



Include inferred: ☑ Results count: 1 (Done in 1,052 sec)

location
<http://www.semanticweb.org/ewallace/ontologies/2019/9/cte_ont4_xfr_events#geospatialRegion-1>

Figure 12. The results of the executed query

Wallace, Evan K.; Ameri, Farhad; Yoder, Reid. "Enabling Traceability in Agri-Food Supply Chains Using an Ontological Approach." Presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC-CIE 2020). August 17, 2020 - August 19, 2020.

CLOSING REMARKS

This paper reports an ongoing effort related to evaluating ontologies and semantic tools and technologies for addressing traceability problems in the agri-food sector. Although an ontology is being developed for the agri-food sector, the underlying traceability model can be applied to all types of manufacturing supply chains in which bulk materials are used. The current version of the ontology can mainly support Transfer and Transport Events and needs to be extended to cover other CTEs including transformation, ownership or custody change, and location change.

In this work, only one external dataset was used to test the triplication process. In future, we are planning to explore new use cases and sample data from AgGateway, use multiple external datasets with varying schema and syntax, and verify the expressivity of the ontology for accommodating disparate datasets.

The SCT Ontology is not sufficiently axiomatized to enable full-scale automated reasoning and inference. Apart from interoperability, ontologies can play a vital role in enabling effective traceability through filling the gaps in incomplete data and reconstructing the otherwise unknown relationships between data entities. For this reason, adding the necessary axioms, without compromising computational efficiency, is the next step in further enhancing the ontology.

Since development of SCRO is still in a work-in-progress, we use the current use case to test the coverage and expressivity of this reference ontology as it is being further enriched and enhanced. Although the traceability ontology can be viewed as an Application Ontology (AO), we expect some portion of the ontology to be merged with the reference ontology.

DISCLAIMER AND ACKNOWLEDGEMENT

Certain commercial systems and applications identified in this paper are not intended to imply recommendation or endorsement by the National Institute of Standards and Technologies, nor is it intended to imply that they are necessarily the best available for the purpose. We acknowledge the input we received from the IOF community and AgGateway subject matter experts. Funding for this research is provided by the National Institute of Standards and Technology (NIST) under collaborative agreement 70NANB19H093.

REFERENCES

[1] Zhang, J., and Bhatt, T., 2014, "A Guidance Document on the Best Practices in Food Traceability," Comprehensive Reviews in Food Science and Food Safety, 13(5), pp. 1074-1103.
[2] Kulvatunyou, B., Wallace, E., Kiritsis, D., Smith, B., and Will, C., 2018, "The Industrial Ontologies Foundry Proof-of-Concept Project," IFIP WG 5.7 International Conference, APMS 2018, Springer, Seoul, South Korea.
[3] 2019, "IOF Charter," https://www.industrialontologies.org/iof-charter/.
[4] Arp, R., Smith, B., and Spear, A. D., 2015, Building Ontologies with Basic Formal Ontology, The MIT Press.
[5] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, R., Schneider, L., and Partner Istc-cnr, L., 2002, WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology.
[6] Gruninger, M., and Menzel, C., 2003, "The Process Specification Language (PSL) Theory and Applications," AI Magazine, 3(24).
[7] Niles, I., and Pease, A., 2001, "Towards a standard upper ontology," Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001, ACM, Ogunquit, Maine, USA, pp. 2-9.
[8] Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V., 2015, "The role of ontologies in biological and biomedical research: a functional perspective," Briefings in Bioinformatics, 16(6), pp. 1069-1080.
[9] Riddick, F. H., Evan K. Wallace , Scott Nieman, Tevis, J., and Ferreyra, R. A., " Implementing Grain Traceability Standards: CART and Simulation," Proc. 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting.
[10] Ameri, F., and Kulvatunyou, B., "Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology," Proc. ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering ConferenceV001T02A052.
[11] Uschold, M., and Gruninger, M., 1996, "- Ontologies: Principles, methods and applications," KNOWLEDGE ENGINEERING REVIEW, 11, pp. 93--136.
[12] 2019, "IOF Technical Principles Document," https://www.industrialontologies.org/iof-technical-principles-document/.
[13] Ceusters, W., and Smith, B., 2015, "Aboutness: Towards Foundations for the Information Artifact Ontology," Proceedings of the Sixth International Conference on Biomedical Ontology (ICBO), CEUR vol. 1515, pp. 1-5.
[14] Stellato, A., Turbati, A., Fiorelli, M., Lorenzetti, T., Costetchi, E., Laaboudi, C., Van Gemert, W., and Keizer, J., "Towards VocBench 3: pushing collaborative development of thesauri and ontologies further beyond," Proc. 17th European Networked Knowledge Organization Systems Workshop, NKOS 2017, CEUR-WS, pp. 39-52.

Deep Learning for Detecting Network Attacks: An End-to-end Approach

Qingtian Zou[1], Anoop Singhal[2], Xiaoyan Sun[3], and Peng Liu[1]

[1] The Pennsylvania State University
{qzz32,pxl20}@psu.edu
[2] National Institute of Standards and Technology
anoop.singhal@nist.gov
[3] California State University, Sacramento
xiaoyan.sun@csus.edu

Abstract. Network attack is still a major security concern for organizations worldwide. Recently, researchers have started to apply neural networks to detect network attacks by leveraging network traffic data. However, public network data sets have major drawbacks such as limited data sample variations and unbalanced data with respect to malicious and benign samples. In this paper, we present a new end-to-end approach to automatically generate high-quality network data using protocol fuzzing, and train the deep learning models using the fuzzed data to detect the network attacks that exploit the logic flaws within the network protocols. Our findings show that fuzzing generates data samples that cover real-world data and deep learning models trained with fuzzed data can successfully detect real network attacks.

Keywords: Network attack, Protocol fuzzing, Deep learning.

1 Introduction

Cyberattacks happen constantly with growing complexity and volume. As one of the most prevalent ways to compromise enterprise networks, network attack remains a prominent security concern. It can lead to serious consequences such as large-scale data breaches, system infection, and integrity degradation, particularly when network attacks are employed in attack strategies such as advanced persistent threats (APT) [13, 26]. Among the different types of network attacks, the *logic-flaw-exploiting network attacks*, which exploit the logic flaws within the protocol specifications or implementations, are very commonly seen. Detecting logic-flaw-exploiting network attacks is very important considering their common presence in APT campaigns. However, it is still a very challenging problem.

Network attack detection methods can mainly be classified into two categories: *host-independent* methods and *host-dependent* methods. The former solely relies on the network traffic, while the latter [4, 10, 21] depends on additional data collected on the victim hosts. The host-dependent methods have some evident drawbacks: they have fairly high deployment costs and operation costs; they are error-prone due to necessary manual configuration by human administrators. Therefore, host-independent detection methods are highly desired as they can decrease deployment and operation costs while reducing the attack surface of

Zou, Qingtian; Singhal, Anoop; Sun, Xiaoyan; Liu, Peng. ”Deep Learning for Detecting Network Attacks: An End-to-End Approach.” Presented at 35TH Annual Conference on Data and Applications Security and Privacy (DBSEC2021). July 19, 2021 - July 20, 2021.

detection system. Unfortunately, we found that the existing host-independent methods, including the classical intrusion detection approaches, often fall short in detecting some well-known and commonly used network attacks.

Recently there is a trend for using machine learning (ML) and deep learning (DL) techniques to detect network attacks. Nevertheless, the DL approaches could also achieve mixed results [8, 14], if they do not address the following two challenges. The first challenge is *useful data sets*. Neural networks require high-quality data and correct labels, which are hard to obtain in real world. Real-world network traffic is often flooded with benign packets, which makes labeling very difficult. Although public data sets [1, 7, 16, 19, 20] for network attacks are available, they are barely useful in detecting logic-flaw-exploiting network attacks due to unbalancing and different focuses. The second challenge is to *identify appropriate neural networks and train the models*. There are a variety of neural network architectures, including multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), etc, which have different characteristics and capabilities. Questions such as which architecture works best for network attack detection, and how to tune the hyper-parameters within models for optimization, are not yet answered.

In this paper, we propose an end-to-end approach to detect the logic-flaw-exploiting network attacks. The end-to-end approach means it starts with acquiring data and ends with detecting attacks using the trained neural networks. To address the data generation challenge, we propose a new protocol fuzzing-based approach to generate the network traffic data. With protocol fuzzing, a large variety of *malicious* network packets for a chosen network attack can be generated at a fast speed. Since the network packets are all generated from the chosen network attacks, they can be labeled as malicious packets automatically without much human efforts. Protocol fuzzing can also generate data with more variations than real world data, or even data that are not yet observed in real world. Moreover, these merits remain when protocol fuzzing is leveraged to generate the needed *benign* network packets. It should be noted that our method is different from data synthesis. Data synthesis is to enhance existing data [11], while our method is to generate new data.

To address the neural network model training challenge, we propose the following procedures: 1) For network attacks (PtH) where we can identify fields of interest, we directly examine the data, and then propose the suitable data representation and neural network architecture. 2) For other network attacks that the field of interests are not obvious, such as DNS cache poisoning and ARP poisoning attacks, we apply different neural network architectures to find out the ones with best performance. We propose to use accuracy, F1 score, detection rate, and false positive rate as the metrics to evaluate the neural networks. All models are trained on the data set with fuzzing involved. We then select the models that work best and evaluate them further on both the fuzzing data set and real attack data set with no fuzzing involved.

The main contributions of this work include: 1) Proposing a DL based end-to-end approach to detect the logic-flaw-exploiting network attacks; 2) Proposing protocol fuzzing to automatically generate high-quality network traffic data for applying DL techniques; 3) Proposing and evaluating neural network models for logic-flaw-exploiting network attack detection; 4) Demonstrating the effectiveness of our approach with three classical logic-flaw-exploiting network attacks, including PtH attack, DNS cache poisoning attack, and ARP poisoning attack.

2 Related Work

The research community has been tackling the network attack detection problem from different perspectives with both classical and novel approaches.

Traditional network attack detection approaches. Traditionally, people usually detect network attacks with approaches such as signature-based, rule-based, and anomaly detection-based methods. In the past, signature-based intrusion detection system (IDS) usually manually crafted signatures [22], which heavily depends on manual efforts. The current techniques focus more on automatic generation of signatures [12]. However, signatures need to be constantly updated to align with new attacks and signature-based detection can be easily evaded by slightly changing the attack payload. Similar problems also exist for rule-based methods [6], which constantly need updates to the rules. As for anomaly detection-based methods, although they require much less manual efforts for updating, they tend to raise too many false positives [3].

Traditional ML and DL for network attack detection. Network attacks are essential for APTs. Some common network attack types include probing, DoS, Remote-to-local, etc. Both traditional ML and DL methods have been adopted for network attack detection. Some focus on one type of network attack and perform binary classifications. For example, MADE [18] employs ML to detect malware C&C network traffic, Ongun et al. [17] employs ML to detect botnet traffic, and DeepDefense [24] employs DL to detect distributed DoS (DDoS) attacks. Others [8, 14, 15, 23, 25] try multi-class classifications, which include one benign class and multiple malicious classes for different kinds of network attacks. The above-mentioned research works all use public data sets.

Network data sets for training and testing detection models. To apply DL for network attack detection, a data set is required. Commonly used public data sets include KDD99 [19], NSL-KDD [7], UNSW-NB15 [16], CI-CIDS2017 [20], and CSE-CIC-IDS2018 [1]. The public data sets are all generated in test-bed environments, with simulated benign and malicious activities. These data sets are often unbalanced due to overwhelming amount of benign data. Even for only malicious activities, multiple types of attacks may be included and the amount of malicious data for each attack type varies a lot. However, balanced data set is important for DL and it is very difficult to label the unbalanced data. Moreover, these data sets focus on network attacks that do not exploit logic flaws, such as DDoS, worms, and C&C over HTTP/HTTPS. They do not contain data about PtH attack, DNS cache poisoning attack, and ARP poisoning attack, which are our detection targets. Though related protocols (e.g. DNS and ARP) are included in those public data sets, such network packets are generated as side effects of other activities, but not because those data sets intentionally want to include such data or launching attacks using those protocols. In a word, existing public data sets are useless in our work.

Protocol fuzzing. Fuzzing is originally a black-box software testing technique, which reveals implementation bugs by feeding mutated data. A key function of fuzzers is to generate randomized data which still follows the original semantics. There are tools for building flexible and security-oriented network protocol fuzzers, such as SNOOZE [5]. Network protocol fuzzing frameworks such as AutoFuzz [2, 9] were also presented. They either act as clients, constructing packets from the beginning, or act as proxies, modifying packets on the fly. We use protocol fuzzing for a different purpose to directly generate high-quality data sets for training neural networks. Instead of using the tools/frameworks mentioned earlier, we prepare our own fuzzing scripts for this specific purpose.

3 Experiment Setup

Since the available public data sets are barely useful for detecting the logic-flaw-exploiting network attacks, this paper will generate comprehensive data sets from scratch, including benign and malicious data sets. We have performed data generation for all three demonstration attacks including PtH, DNS cache poisoning, and ARP poisoning. ARP poisoning attack only requires one malicious packet for a successful attack, so we call it the single-packet attack. PtH and DNS cache poisoning attacks, however, need multiple malicious packets for one successful attack, so we call them multi-packet attacks. Due to page limits, we only discuss data generation details about multiple-packet attacks in this section because they are more complicated than single-packet attacks. Below subsections discuss the general approach and implementation principles of protocol fuzzing followed by attack-specific details. All attacks are carried out thousands of times so that a fair amount of malicious data can be collected. Benign data generation also lasts long enough to gather the commensurate amount of data compared to malicious data. The network packet capturing is performed at the victim's side. After that, information about detection neural networks are provided.

3.1 Protocol Fuzzing and The Implementation

In client-server enterprise computing, the server-side protocol implementations are often complex and error-prone. Hence, there is a need to achieve thorough testing of the server-side implementation. Protocol fuzzing tools [2, 5, 9] are usually functioning at the client side, so that unexpected errors on the tested server programs may be triggered. A main difference between protocol fuzzing and software fuzzing is that the protocol specification, especially its state transition diagram, will be used to guide the fuzzing process. In this way, fuzzing tests could be performed in a stateful manner.

This paper leverages protocol fuzzing to change the contents of network packets, specifically, the values of some fields in the packets. If a field is to be fuzzed, it will be assigned with pre-determined values, rather than the values chosen by the network client program. The fuzz fields are chosen based on the following steps: 1) All fields in the packet of the attack-specific protocol are considered. 2) One field on the list will be picked and fuzzed by assigning pre-determined values, rather than values that are normally provided by the network programs. 3) The success rate of the attack after fuzzing the field will be monitored. If the attack success rate is above 50%, it confirms that this field can be fuzzed. 4) After one field is fuzzed, the above steps will be repeated for the next field on the list, while keeping the already fuzzed field(s) still fuzzed.

```
Result: BList, which stores fields to fuzz
input AList of all available fields;
initialize an empty BList to store fields to fuzz;
foreach field in AList do
    fuzz field;
    fuzz all fields in BList;
    launch the attack for hundreds of times;
    count successful attacks and calculate success rate;
    if attack success rate is over 50% then
        add field to BList;
    end
end
```

Algorithm 1: Select fields to be fuzzed.

To ensure the fuzzed packets are valid, we need to firstly make sure *AList*, the list of all candidate fuzzing fields, does not contain fields that will affect the packets' integrity, such as fields of checksum values and packet lengths. The values of those fields should not be arbitrarily changed. Furthermore, when we choose the fields to be fuzzed, we need to make sure the attack success rate after fuzzing this field is always above 50%.

An additional benefit of protocol fuzzing is that it can generate and cover malicious data samples which may otherwise be overlooked when applying deep learning. In deep learning, the changed values for the fuzzing fields may make the malicious data samples misclassified as benign. With protocol fuzzing, if the malicious data are generated in attacks, they'll be labeled as malicious automatically. Thus, these malicious data samples won't be omitted in the malicious data set.

3.2 PtH

PtH Attack. PtH 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 whether hashes or the calculation results of them matches or not. PtH relies on these vulnerable mechanisms to impersonate normal users with dumped hashes. We assume that: (a) normal users use benign client programs that are usually authenticated through more reliable mechanisms other than just using hashes, 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 at the server side and find out which kind of authentication mechanism is used by a user: the more reliable mechanism, or the vulnerable mechanism using only hashes. The login sessions using those vulnerable authentication mechanisms can then be identified as PtH attack.

Windows remote login processes, if not properly configured, can use such vulnerable authentication mechanisms. Windows remote login can be divided

Table 1: Fields of interest.

Layer	Fields	Size in bytes	Explanation
ETH	Dst_MAC	6	Destination MAC address
	Src_MAC	6	Source MAC address
	ETH_type	2	Indicate which protocol is encapsulated in the payload of the frame.
ARP	HTYPE	2	Network link protocol type. For Ethernet, this field is 1.
	PTYPE	2	For IPv4, this value should always be 0x0800.
	HLEN	1	Length of a hardware address. For Ethernet addresses, the length is 6.
	PLEN	1	For IPv4 addresses, this value should always be 4.
	OpCode	2	Specifies the operation that the sender is performing: 1 for request, 2 for reply.
	SHA	6	Source hardware (MAC) address.
	SPA	4	Source internetwork (IP) address.
	THA	6	Target hardware (MAC) address.
	TPA	4	Target internetwork (IP) address.
IP	Version	4/8	For IPv4, this is always equal to 4.
	IHL	4/8	Internet header length.
	DSF	1	Differentiated service field, which includes differentiated services code point and explicit congestion notification.
	TLen	2	The entire packet size in bytes.
	ID	2	Identification field.
	Flags	3/8	3 bits for controling or identifying fragments.
	FragOff	13/8	Fragment offset.
	TTL	1	Time to live field, which limits a datagram's lifetime.
	prot	1	This field defines the protocol used in the data portion of the IP datagram.
	chksum	2	Header checksum for error-checking of the header.
	src_add	4	Source IPv4 address.
	dst_add	4	Destination IPv4 address.
UDP	src_port	2	Source port number.
	dst_port	2	Destination port number.
	hd_len	2	The length in bytes of the UDP header and UDP data.
	chksum	2	Checksum field for error-checking of the UDP header and UDP data.
DNS	TID	2	Transaction ID.
	flags	2	Control flags
	q	2	The number of entries in the question section.
	AnRR	2	The number of resource records in the answer section.
	AuRR	2	The number of resource records in the authoritative section.
	AdRR	2	The number of resource records in the additional section.
SMB/ SMB2	cmd	2	The command code of this packet.
	flags	4	Indicate how to process the operation
	NT_status	4	Status or error code.

into three stages, protocol and mechanism negotiation (initial communication), authentication, and task-specific communication (afterwards communication). Each stage contains multiple network packets, and hashes are used in the authentication stage for impersonation. The authentication stage 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. 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 hashes 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.

Data generation. We set up a Windows 2012 Server R2 as the victim server machine, a Windows 7 as the user client machine, and another Kali Linux as the attacker machine. The data sets are automatically generated by protocol fuzzing, and the protocol of interest here is Server Message Block (SMB), or a newer version of it, denoted as SMB2. SMB/SMB2 provides functions including file sharing, network browsing, printing, and inter-process communication over a network. In our data generation, more than 15 fields are fuzzed in each SMB/SMB2 packets, including `SMB flags`, `SMB capabilities`, and fields in SMB header, etc. e leverage the PtH script in Metasploit Framework to launch the attack. Boxes connected with solid lines are what happens at foreground, and boxes in the dash line area happen behind the scene. The process is to start the Metasploit Framework, set exploit parameters, start the exploitation, and then wait 25 seconds while monitoring the attack status. If the waiting time is too short, the attack may be stopped before completion. While the console is waiting at the foreground, the exploitation is ongoing at the background. Network packets in all the three stages, initial communication, authentication, and afterwards communication, are fuzzed. After the exploitation, based on whether the attack succeeds or not, we may continue to establish C&C, like what a real attacker will do. (The C&C network traffic are mainly TCP packets, which are not used for attack detection. Details are discussed later.) Finally, we quit all possibly established sessions and the Metasploit Framework, and then either freshly start another fuzzing iteration to generate more data or stop. The sign of a successful PtH attack is an established reverse shell, which can be observed at the attacker's side.

The same fuzzing method has also been applied in the generation of benign data. We first prepare a list of normal commands, including files reading, writing, network interactions, etc. For each benign fuzzing iteration, we randomly choose a command from the list, and then use valid username, plain-text password, and tool to log in to the server and execute the command.

All the network packets from malicious and benign network traffic are captured using Wireshark at the victim's side. Due to fuzzing, not all PtH attempts or benign access attempts can be guaranteed to succeed. For failed PtH attempts, we remove them from malicious data because they do not generate real malicious impact, and they cannot be categorized as benign either because they are generated with attacker tools for malicious purpose. For failed benign accesses, we keep them in benign data, because normal users can also have failed logins due to typos, wrong passwords, etc.

In one PtH attack, there are packets for initial communications, authentication and afterwards communications. One data sample consists of multiple packets, and those packets may come from one, two, or all of the three stages above. Besides, one complete PtH attack or benign activity most certainly contains more packets than one data sample can represent. When labeling, if the session is malicious, then all data samples generated from this session is labeled malicious, and the same is also true for the benign cases.

Detections. To detect PtH attack with neural networks, we have two key insights that help determine the representation of data samples: 1) Network communication for authentication is actually a sequence of network packets in certain order. An earlier packet can affect the packet afterwards. For example, the first several packets between a server and a client may be used to communicate and determine which protocol to use (e.g. SMB or SMB2), and packets afterwards will use the decided protocol. The attack is to get authenticated by the server, which requires a sequence of packets to accomplish. Therefore, each data sample should be a sequence of packets, rather than an individual packet. 2) PtH relies on authentication mechanisms that legitimate users usually don't use. The network packets for the benign and malicious authentication are different. Since both authentication methods use SMB/SMB2 packets, the differences between them thus exist in the fields of the SMB/SMB2 layer. Therefore, data in `SMB/SMB2` layer is used for PtH detection. In addition, the differences of field values between benign and malicious authentication will be helpful to distinguish them. For this attack, we choose Long-short term memory (LSTM) as the architecture for the neural network.

3.3 DNS Cache Poisoning

DNS cache poisoning. A major functionality of DNS is to provide the mapping between the domain names and IP addresses. When a client program refers to a domain name, the domain name needs to be translated to an IP address. The DNS servers are responsible to perform such translation.

The global DNS system has a hierarchical structure that contains root name servers, top-level domain name servers, and authoritative name servers. Some examples are the public DNS servers `8.8.8.8` and `8.8.4.4` provided by Google, and recently released `1.1.1.1` by Cloudflare. These name servers, referred as the global DNS servers, provide records that maps the domain names and IP addresses. Due to the geological distance between user machines and the global DNS servers, it is very costly to contact the global DNS servers every time very often. To reduce the cost, organizations deploy their own DNS servers, referred as local DNS servers, within the LAN to cache the most commonly used mappings between domain names and IP addresses. Generally, when a user machine needs to make connection with a destination machine, it will contact the local DNS server first to resolve the domain name. If the local DNS server does not cache the DNS record for this domain name, it will send out a DNS query to the global DNS server to get the answer for the user machine. The user machine gets to know the IP address after receiving the response.

DNS cache poisoning attack can target local DNS servers. When the local DNS server receives a query which it does not have the corresponding records (first stage), it will inquire the global DNS server (second stage). On receiving the response (third stage), the local DNS server saves this record in its cache to avoid inquiring the global DNS again when receiving the same query. It then forwards the response to the user machine (fourth stage). However, the DNS server cannot verify the response at the third stage, and this is where the attacker can fool the local DNS server. Pretending as the global DNS server, the attacker can send a spoofed DNS response to the local DNS server with falsified DNS records. If the fake response arrives earlier than the real one, the local DNS server will save the falsified record to its cache and forward it to the user machine. When new queries about the same domain name comes in, the local DNS server will not query the global DNS server again because the corresponding record has been cached. Consequently, it will answer the user machine with the falsified record, until the record expires or the cache is flushed.

Data generation. For this attack, ten fields, such as `time to live` values in different layers, are fuzzed. The test bed contains three machines: a local DNS server whose DNS cache is flushed periodically, a user machine which sends out DNS queries to the local DNS server periodically, and an attacker machine which sniffs for DNS requests sent by the local DNS server and answers them with spoofed responses as in the attack scenario, or does nothing otherwise.

In the malicious scenario, we make the user machine ask for the IP address of one specific domain name from the local DNS server using command *dig*. The domain name is one that does not have a corresponding record on the local DNS server, thus enabling the DNS cache poisoning attack towards it. The attacker machine sniffs for DNS queries with that specific domain name sent out from the local DNS server, and responds them with fuzzed DNS responses with falsified IP addresses. Then the DNS cache gets poisoned and the user machine gets the falsified DNS record. We keep the user machine sending out DNS queries periodically, so that the above process repeats many times and a large amount of data can be generated. However, as discussed earlier, if the local DNS server has the record for the domain name in its cache, it will not send out DNS queries for it. This is why we flush the DNS cache of the local DNS server, so that it remains vulnerable in different iterations. If the attack is successful, the falsified IP addresses can be seen on the results of *dig*.

In the benign scenario, we prepare a list containing 4098 domain names. In each iteration, the user machine randomly chooses one domain name from the list, and sends a request to the local DNS server. To resemble the malicious scenario, the cache of local DNS server is also flushed periodically so that the local DNS server always needs to communicate with the global DNS server.

The domain name used in the malicious scenario and the domain names used in the benign scenario do not overlap. Both the domain names and the IP addresses (falsified or genuine) are excluded during training, which can be treated as signatures. Because DNS cache poisoning is a multi-packet attack, the labeling to data samples is also based on sessions, similar to PtH attack.

Detections. Network packets from DNS cache poisoning attack form sessions which consist of queries and answers. Therefore, each data sample should include data from multiple network packets. In addition, it is not clear which fields may be of importance, so we need to investigate the packet content, rather than simply generalizing the packets with packet types as we did in PtH detection. The data samples are processed to be image-like. That is, each row represent one packet, and each element in the row represent one byte in that packet. We use a convolutional neural network (CNN) to do the classifications, which has been proven to work well in image classification problems. The labeling is done towards each data sample, which is the entire matrix, rather than an individual packet. During the data processing process, the malicious and benign data are processed separately. Matrices generated from malicious data are labeled as malicious, and matrices from benign data are labeled as benign. Similar to PtH detection, we have trained a series of neural networks with different neural network hyper-parameters and data samples of different window sizes and window steps. That means we can adjust the number of packets k included in each data sample and thus change the size of matrix.

4 Evaluations

This section provides the evaluation results of the three demonstration attacks on the selected best-performing and best-detecting models. For comparison with DL models, we have also trained traditional ML models, including k-nearest neighbor (kNN) models, support vector machine (SVM) models with various kernels, decision tree (DT) models, and random forest (RF) models. They are trained, selected, and evaluated on the same data sets. For PtH and ARP poisoning, the traditional ML models' data samples and features are the same as those for DL models. However, for DNS cache poisoning, the same data sample and feature cannot be used because the input space is too large for traditional ML models to handle. Therefore, we employed principal component analysis (PCA) for dimension reduction, and only select the top-rated one-fifth PCA features. On average, they can explain about 97.09% of the original data.

4.1 Model Selection

For model selection, we consider not only the perspective of neural network performance, but also the perspective of security. We use accuracy (*Acc* and F1 score (*F1*), two commonly used metrics, to measure the classification, and use detection rate (*DR*) and false positive rate (*FPR*) for attack detection effectiveness. Assuming the numbers of true positives, true negatives, false positives and false negatives are presented as TP, TN, FP, FN, respectively, then $Acc = (TP+TN)/(TP+TN+FP+FN), F1 = TP/(TP+0.5*(FP+FN)), DR = TP/(TP+FN)$, and $FPR = FP/(TN+FP)$. *DR* shows the detector's ability of detecting attacks. *FPR* shows how likely the detector raises false alarms. We call the best-performing model as the one that gets the highest average of *Acc* and *F1*, denoted as $P = \frac{Acc+F1}{2}$, and the best-detecting model as the one that gets the highest average of *DR* and $1-FPR$, denoted as $D = \frac{DR+1-FPR}{2}$. If FPR cannot be calculated (no benign data sample), we let $D = DR$. We simply take the average because all the chosen metrics are equally important for evaluations.

The generated fuzzing data set is randomly split into two parts: 80% as the training set, and 20% as the test set. The training set is then further randomly split into four parts of about the same size, upon which 4-fold cross-validation is employed to avoid over-fitting. All the reported results are the average results among four folds. The best-performing and best-detecting models are selected based on the average P and D results on the validation set across all four folds.

4.2 Data Sets

Table 2 shows the data set statistics. The data set contains fuzzed set (split into training set and test set) and non-fuzzed set (real attack set). A data set with sufficient and balanced data samples is essential for training the models effectively. Lack of training data can result in poor results, while biased data sets may result in biased models. If the fuzzing data set is already balanced, we directly use all the data samples without balancing. Otherwise, we perform data set balancing first. Specifically, if the benign data sets have significantly more data samples than the malicious data sets, we down-sample the benign data sets to match the size of malicious data sets, and vice versa.

Table 2: Data set statistics.

Attacks	Set	Size	Benign to malicious ratio
ARP poisoning	Training	9584	1.005:1
	Test	2400	0.982:1
	Real attack	17471	0:1
PtH*	Training	3932	1.364:1
(best-performing)	Test	983	1.329:1
	Real attack	214	0:1
PtH*	Training	2556	0.974:1
(best-detecting)	Test	640	0.839:1
	Real attack	192	0:1
DNS	Training	30928	1.003:1
cache poisoning*	Test	7732	0.988:1
	Real attack	263	0:1

* For multi-packet attacks, we only list the data set statistics corresponding to the best-performing or best-detecting models.

4.3 Best-performing Models

Table 3 presents the evaluation results on the best-performing models for each network attack. All models get acceptable to good results on training set and test set. **For multi-packet attacks, DL models are substantially better than traditional ML models, especially on real attack set.** In PtH detection, the LSTM model achieves near 99% accuracy on the real attack set, while ML models cannot reach 1/4 accuracy. In DNS cache poisoning detection, the CNN model's accuracy on the real attack set is 100%, while ML model can reach about 47% accuracy at most. Selected DL models' F1 scores are also far better than those of traditional ML models. For ARP poisoning detection, DL models do not have many advantages over traditional ML models, and all models' performances downgrade on real attack set comparing to those of training set and test set. The reason is that the real attack set for ARP poisoning is generated on a different LAN, with different valid MAC and IP addresses.

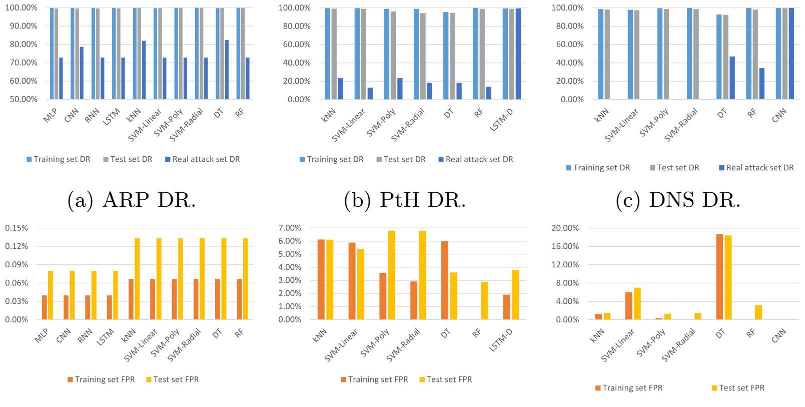
Table 3: Evaluation results on best-performing models.

Attacks	DL or ML	Model type1	Training set		Test set		Real attack set	
ARP	DL	MLP	99.91%	0.9991	99.75%	0.9975	72.84%	0.8429
		CNN	99.94%	0.9994	99.79%	0.9979	73.02%	0.8441
		RNN	99.91%	0.9991	99.75%	0.9975	72.83%	0.8428
		LSTM	99.91%	0.9991	99.75%	0.9975	72.83%	0.8428
	ML	kNN	99.90%	0.9990	99.93%	0.9993	81.99%	0.9010
		SVM-Linear	99.87%	0.9987	99.90%	0.9990	72.83%	0.8428
		SVM-Poly	99.96%	0.9996	99.93%	0.9993	72.83%	0.8428
		SVM-Radial	99.97%	0.9997	99.93%	0.9993	72.83%	0.8428
		DT	99.84%	0.9984	99.90%	0.9990	82.35%	0.9032
		RF	99.97%	0.9997	99.93%	0.9993	72.83%	0.8428
PtH	DL	LSTM-P	98.45%	0.9865	98.07%	0.9831	98.96%	0.9948
	ML	kNN	96.77%	0.9682	96.53%	0.9658	23.44%	0.3797
		SVM-Linear	96.89%	0.9694	96.72%	0.9674	13.02%	0.2304
		SVM-Poly	97.75%	0.9779	94.69%	0.9479	23.44%	0.3797
		SVM-Radial	98.07%	0.9810	93.72%	0.9378	18.23%	0.3084
		DT	94.70%	0.9467	95.44%	0.9533	18.23%	0.3084
		RF	100.00%	1.0000	97.99%	0.9798	14.06%	0.2466
DNS	DL	CNN	99.87%	0.9987	99.73%	0.9973	100.00%	1.0000
	ML	kNN	98.67%	0.9867	98.35%	0.9834	0.00%	0.0000
		SVM-Linear	96.01%	0.9608	95.17%	0.9527	0.00%	0.0000
		SVM-Poly	99.63%	0.9963	98.70%	0.9870	0.00%	0.0000
		SVM-Radial	100.00%	1.0000	98.66%	0.9867	0.00%	0.0000
		DT	87.01%	0.8771	86.88%	0.8754	47.01%	0.6395
		RF	100.00%	1.0000	97.50%	0.9752	34.19%	0.5096

[1] For multi-packet attacks, only proposed DL models are presented.

4.4 Best-detecting Models

Figure 1 presents the evaluation results of best-detecting models. FPRs on real attack sets are not presented because there is no negative data sample, so FPR



(a) ARP DR. (b) PtH DR. (c) DNS DR.

(d) ARP FPR. (e) PtH FPR. (f) DNS FPR.

Fig. 1: Evaluation results on the best-detecting models.

cannot be calculated. Similar to the best-performing case, all models get acceptable to good results on training and test set. For single-packet attack detection, DL models do not have many advantages over ML models. **For multi-packet attacks, DL models are better than ML models, especially on real attack set.** Because there is no negative data sample in the real attack set, $DR = Acc$. As for FPR, although it cannot be calculated in the real attack set, results show that DL models achieve generally lower FPRs comparing to ML models on the training and test sets.

5 Discussions and Limitations

Lack of efficiency: Training a neural network requires a large amount of data samples. However, the number of data samples can be affected in many ways. For example, protocol fuzzing in nature cannot guarantee that all malicious/benign activities are successful. Although the fuzzed values are in a valid range, the network packets with fuzzed values may still get rejected by the server or trigger some unexpected circumstances, leading to an interrupted session. Those data are probably useless as discussed in section 3.2. Also, the removal of duplicate (same data in one class) and double-dipping (same data among different classes) data samples will also affect the number of data samples. In a word, not all collected data can be used as data sample for neural network training.

Another factor that affects the efficiency is the time consumed by each benign/malicious activity. Except for some simple activities like MAC-IP address resolving with only several ARP packets, other complicated activities need time to carry out, especially those containing hundreds or more network packets. Moreover, depending on the mechanism of packet processing, the client/server may also need more time before it can respond. For example, in PtH data generation, one successful attack contains 300 to 400 packets (and not all of them are usable to generate data sample), and some time intervals between adjacent packets can be as large as 0.5 second. In addition, in our experiments, we manually inserted idle time intervals. This time interval is reserved so that the exploitation can continue to run to reach a successful end. If this time interval is removed or too short, then the attack process is very likely to end in the middle of exploitation. In a word, each data generation iteration takes time to complete.

As a result, our data generating efficiency is not very high. Take PtH as an example, we spent about 4 days running 5,000 attack iterations, of which 611 failed. The total amount of network packets captured is 497,956, of which 103,718 are related packets. However, the final number of data samples is only in the thousands, as shown in Table 2.

Neural networks for various network attacks: Though we have verified our idea on three chosen network attacks, we trained separate neural networks for different attacks. We can not train a generic neural network to detect various network attacks. It is difficult to train such a neural network because different network attacks have different characteristics, which may need different data representations and neural network architectures.

Impact of probability threshold: The raw outputs for output layers of the detection neural networks are the probabilities for the data sample to be benign or malicious, which add up to 1. The raw outputs can be converted to classification results. If the probability for malicious class is beyond a threshold (e.g., 0.5), then the data sample is classified as malicious. Otherwise, it is classified as benign. When the probability threshold increases, the model is more likely to classify a data sample as benign, and thus decrease detection rates and false positive rates. The probability threshold can be tuned depending on whether the defender prefers higher detection rates or lower false positive rates.

6 Conclusion

This paper presents an end-to-end approach to detect the logic-flaw-exploiting network attacks using DL. The end-to-end approach begins with data generation and collection, and ends with attack detection with neural networks. We address two major challenges in applying DL for logic-flaw-exploiting network attack detection: the generation of useful data sets and the training of appropriate neural network models. We show the effectiveness of our approach with three specific demonstration attacks, including PtH, DNS cache poisoning, and ARP poisoning. We have generated high quality network traffic data using protocol fuzzing, trained neural networks with generated data, and evaluated the trained models from the perspective of both neural network performance and attack detection. We have also discussed the limitations of our experiments and approach.

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. IDS 2018 | Datasets | Research | Canadian Institute for Cybersecurity | UNB (Jan 2020), https://www.unb.ca/cic/datasets/ids-2018.html, [Accessed Jul 4 2020]

2. Aitel, D.: The advantages of block-based protocol analysis for security testing. Immunity Inc., February **105**, 106 (2002), `http://www.immunityinc.com/downloads/advantages_of_block_based_analysis.pdf`
3. Amini, M.e.a.: Rt-unnid: A practical solution to real-time network-based intrusion detection using unsupervised neural networks. computers & security **25**(6), 459–468 (2006)
4. Arote, P., Arya, K.V.: Detection and prevention against arp poisoning attack using modified icmp and voting. In: 2015 International Conference on Computational Intelligence and Networks. IEEE (2015)
5. Banks, G.e.a.: SNOOZE: Toward a Stateful NetwOrk prOtocol fuzZEr. In: Springer, pp. 343–358 (2006)
6. Choi, J.e.a.: A method of ddos attack detection using http packet pattern and rule engine in cloud computing environment. Soft Computing **18**(9), 1697–1703 (2014)
7. Dhanabal, L., Shantharajah, S.: A study on nsl-kdd dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering **4**(6), 446–452 (2015)
8. Faker, O., Dogdu, E.: Intrusion detection using big data and deep learning techniques. In: ACMSE 2019 - Proceedings of the 2019 ACM Southeast Conference (2019)
9. Gorbunov, S., Rosenbloom, A.: AutoFuzz: Automated Network Protocol Fuzzing Framework. International Journal of Computer Science and Network Security **10**(8), 239–245 (2010)
10. Goswami, S.e.a.: An unsupervised method for detection of xss attack. IJ Network Security **19**(5), 761–775 (2017)
11. Jan, S.T.e.a.: Throwing darts in the dark? detecting bots with limited data using neural data augmentation. In: The 41st IEEE Symposium on Security and Privacy (IEEE SP) (2020)
12. Kaur, S., Singh, M.: Automatic attack signature generation systems: A review. IEEE Security & Privacy **11**(6), 54–61 (2013)
13. Milajerdi, S.M.e.a.: Holmes: Real-time apt detection through correlationof suspicious information flows. In: 2019 IEEE Symposium on Security and Privacy (SP). IEEE (2019)
14. Millar, K.e.a.: Deep learning for classifying malicious network traffic. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 11154 LNAI (2018)
15. Mishra, P.e.a.: A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys Tutorials **21**(1), 686–728 (2019). https://doi.org/10.1109/COMST.2018.2847722
16. Moustafa, N., Slay, J.: UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings. Institute of Electrical and Electronics Engineers Inc. (dec 2015)
17. Ongun, T.e.a.: On designing machine learning models for malicious network traffic classification. arXiv:1907.04846 [cs, stat] (Jul 2019), `http://arxiv.org/abs/1907.04846`, arXiv: 1907.04846
18. Oprea, A.e.a.: Made: Security analytics for enterprise threat detection. In: Proceedings of the 34th Annual Computer Security Applications Conference. ACSAC '18, Association for Computing Machinery (Dec 2018). https://doi.org/10.1145/3274694.3274710
19. Pfahringer, B.: Winning the kdd99 classification cup: bagged boosting. ACM SIGKDD Explorations Newsletter **1**(2), 65–66 (2000)

20. Sharafaldin, I.e.a.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy. vol. 2018-Janua (2018)
21. Sun, H.M.e.a.: Dependns: Dependable mechanism against dns cache poisoning. In: International Conference on Cryptology and Network Security. pp. 174–188. Springer (2009)
22. Taylor, C.e.a.: Low-level network attack recognition: a signature-based approach. IEEE Proc. PDCS'2001 (2001)
23. Yin, C.e.a.: A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. IEEE Access **5**, 21954–21961 (2017)
24. Yuan, X., Li, C., Li, X.: DeepDefense: Identifying DDoS Attack via Deep Learning. In: 2017 IEEE International Conference on Smart Computing, SMARTCOMP 2017 (2017)
25. Zhang, Y.e.a.: PCCN: Parallel Cross Convolutional Neural Network for Abnormal Network Traffic Flows Detection in Multi-class imbalanced Network Traffic Flows. IEEE Access pp. 1–1 (2019)
26. Zou, Q.e.a.: An approach for detection of advanced persistent threat attacks. IEEE Annals of the History of Computing **53**(12), 92–96 (2020)

Study of Multicast Broadcast Single Frequency Network Area in Multicast Communication

Chen Shen*†, Chunmei Liu‡, Richard A. Rouil‡, and Hyeong-Ah Choi§
*Associate, Wireless Networks Division, National Institute of Standards and Technology, USA
†Department of Physics, Georgetown University, USA
‡Wireless Networks Division, National Institute of Standards and Technology, USA
§Department of Computer Science, George Washington University, USA
Email: †sc1951@georgetown.edu ‡{chunmei.liu, richard.rouil}@nist.gov, §hchoi@gwu.edu

***Abstract*—Multicast was introduced in Long-Term Evolution (LTE) in release 8 as the technology for the Evolved Multimedia Broadcast Multicast Services (eMBMS) and has been steadily updated in the later releases. Unlike unicast where available resources are split amongst each User Equipment (UE), in Multicast Broadcast Single Frequency Network (MBSFN), all the resources can be allocated to reach all UEs. This prominent advantage together with the prevalent signal improvement make MBSFN a popular LTE and 5G evolution technique. In this paper, we focus on the LTE Multicast Broadcast Single Frequency Network (MBSFN) technology from the perspective of the MBSFN area and its application in public safety networks (PSNs). Since the MBSFN area plays a pivotal role in performance improvement and public safety incidents differ in their scales, using our state-of-the-art system-level simulations, we explore performance distributions within MBSFN areas of different shapes and sizes, as well as the impacts of MBSFN area locations relative to the network. We show that the results can give an informative insight into the MBSFN area for in-coverage performance prediction/evaluation or out-of-coverage mobile base station deployments.**

***Index Terms*—LTE, MBSFN area, resource efficiency.**

I. Introduction

While the majority of Long-Term Evolution (LTE) downlink transmissions uses unicast to achieve excellent performance, when transmitting the same content to a large group of User Equipment (UEs), LTE Multicast Broadcast Single Frequency Network (MBSFN) shows excellent potential performance improvement over unicast [1]. To serve this type of traffic, in unicast, available resources are split into sets of resource blocks (RBs) and each UE gets its own share. The content is then mapped to each RB set and transmitted to individual UEs separately. Whereas in MBSFN, multiple cells are configured to form an MBSFN area and tightly synchronized. These cells map the content to the entire available resources and transmit identical waveforms together to all the UEs. From UE perspective, this is equivalent to that each individual UE has full access to the entire available resources. Consequently, in theory, MBSFN can serve an unlimited number of UEs simultaneously, and hence, it can boost the spectral efficiency significantly when serving an enormous number of users [2]. For this reason, MBSFN is of particular interest in public safety communications due to their significant group traffic with a multicast nature.

Supporting multicast transmissions in MBSFN has some limitations. First, spatial multiplexing technology, which plays an essential role in high LTE unicast performance, is not supported in MBSFN. Next, the current MBSFN protocol lacks the transmission acknowledgments from each UE. To get reliable transmissions, a more conservative Modulation and Coding Scheme (MCS) or Channel Quality Indicator (CQI) is preferred [3] and is combined with a stricter target Block Error Rate (BLER), which is usually 0.01 compared with 0.1 in unicast [4]. Besides, in early LTE releases, the number of subframes allowed for MBSFN is limited to 6 out of 10. This constraint is relaxed in later releases. Furthermore, due to the larger serving size of the MBSFN area, longer Cyclic Prefix (CP) is required. With more reference signals required, the resulting larger overhead leads to much fewer physical resources available for data in each frame, compared with unicast [5].

Last but not least, all UEs served by MBSFN must use the same MCS, which is typically selected to be the minimum of all UEs' to guarantee proper demodulation and decoding [6]. For instance, if the minimum MCS index of all UEs in one MBSFN area is 8, then MCS 8 will be applied to all despite some UEs being able to use higher MCS. In LTE release 15 Table 7.1.7.1-1 [7], MCS index values are defined from 0 to 31, with higher values having higher spectral efficiency.

Despite the above limitations, MBSFN can still outperform unicast in some cases by allowing each UE full access to the entire available resources and by Signal-to-Interference-plus-Noise Ratio (SINR) improvement. Both factors can hardly be quantified for general cases. The first one, full access to the entire available resources, plays a primary role in MBSFN. The more UEs, the higher the spectral efficiency. The only drawback of increasing the UE size is the potential lower MCS for MBSFN, given that the minimum MCS among UEs will be chosen. The second one, SINR improvement, comes from two sources: less inter-cell interference and constructive signal combination from MBSFN cells, which have been analyzed in detail in [8]. In principle, the UE locations relative to the MBSFN area determine the distribution of useful signals and interference, thus different SINR improvement and the resulting performance across the MBSFN area.

This difference in performance across MBSFN areas is not of particular importance in commercial multicast/broadcast due to their large geographic area and that most UEs are at the center. Hence there are little such studies in the literature.

Shen, Chen; Liu, Chunmei; Rouil, Richard A.; Ah-Choi, Hyeong. "Study of Multicast Broadcast Single Frequency Network Area in Multicast Communication." Presented at 14th International Conference on Signal Processing and Communication Systems (ICSPCS 2020). December 14, 2020 - December 16, 2020.

However, the areas covered by public safety incidents differ in size and shape, leading to significant performance differences.

With this in mind, in this paper, we investigate the impact of the MBSFN area and UE locations on the performance with various MBSFN deployment scenarios by using our cutting edge system-level simulations, which take into account all the MBSFN factors mentioned above [9]. When macro-cells are in place in the public safety incident area, our results can give precise performance prediction; For areas without underlying macro-cells, our results can provide guidelines to deploy mobile base stations to serve public safety traffic for various scenarios [10] [11]. To our knowledge, this work is the first of its kind for the MBSFN areas of different schemes with extensive simulation results.

The rest of the paper is organized as follows. In Section II, we briefly describe our system-level simulation and some necessary configurations for the MBSFN settings. In Section III, we present and analyze our simulation results for different MBSFN deployment scenarios. At last, we summarize our observations in Section IV.

II. System Simulation Configurations for MBSFN

In typical non-Coordinated Multi-Point (non-COMP) unicast transmissions, each UE is served by one cell, and the coordination between neighboring cells is limited mainly for interference reduction and handovers. Whereas in MBSFN transmissions, several MBSFN cells form an MBSFN area where UEs inside the area can be served by all the corresponding MBSFN cells. Since the transmitting signals come from several cells, the UEs' positions relative to the area, together with the MBSFN area size and shape, will influence the performance. To explore this impact, we design multiple scenarios for unicast and MBSFN transmissions with different area sizes and shapes.

First, we set the basic scenario of hexagonally distributed three rings of cells for both unicast and MBSFN, as in Figure 1. The red dots with black triangles represent tilted tri-sector sites, each of which counts for three cells. The outer two rings of cells are only to generate interference, and we are only interested in the performance of the central 21 cells (7 tri-sector sites), colored in grey and blue in Figure 1. The antenna configuration for the downlink is eight transmitters and four receivers for both unicast and MBSFN, and transmit mode (TM) 9 is used for the unicast. The simulation length in both cases is 1 second, which counts for 1000 subframes or TTIs (Transmission Time Interval) in the simulations. We also assume zero feedback delay and perfect channel knowledge for the modeling and simulations.

In unicast, due to the symmetricity of the network layout and to reduce the simulation time, we further shrink the area of focus from 21 cells to 7 cells, colored in blue in Figure 1. The different transparencies in the figure indicate the different cell areas.

The largest MBSFN area in our design consists of the central 21 cells, as shown in Figure 2, where the numbers from 1 to 21 indicate the cell indices. The areas colored with different transparencies, but the same hue represents three cells of one site. For MBSFN, the performance gain mainly comes from three parts: full physical layer resources for all UEs, the signal combination from all MBSFN cells, and the possible interference reduction from neighboring cells. The latter two gains depend on the MBSFN area's topology but are still centrosymmetric in the basic scenario. Thus we only consider UEs in the selected seven cells (Cell IDs 1 to 8 except 7) as discussed before. It can be seen that this setting extensively reduces simulation computations yet covers UEs at the center, middle, and edge all over the MBSFN area.

Figure 3 further illustrates the constructive signals contributed by multiple cells in the MBSFN area, represented

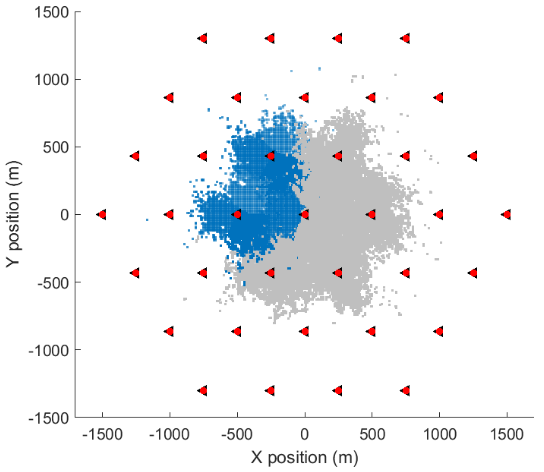


Fig. 1: UE Saturated in 21 and 7 Cells, Respectively

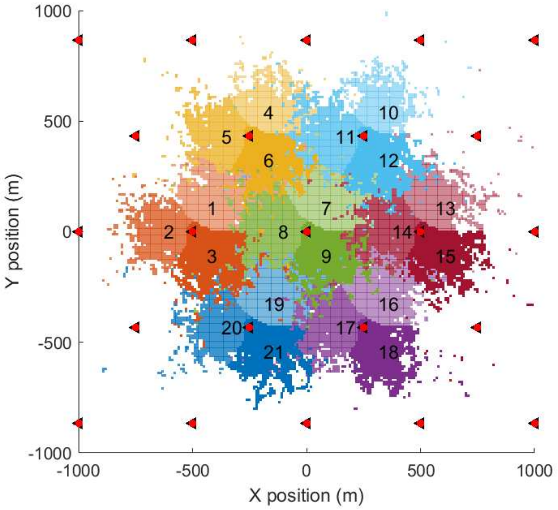


Fig. 2: Network Layout

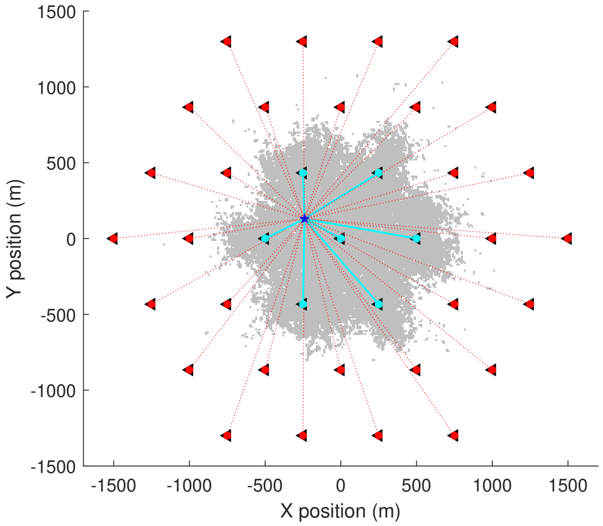


Fig. 3: MBSFN: Signal and Interference

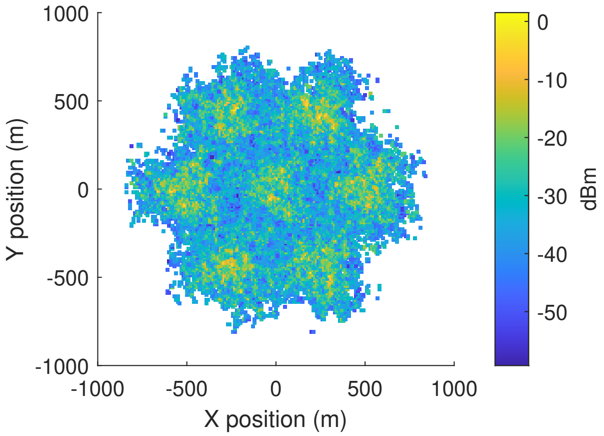


Fig. 4: Unicast: Wide-band Signal

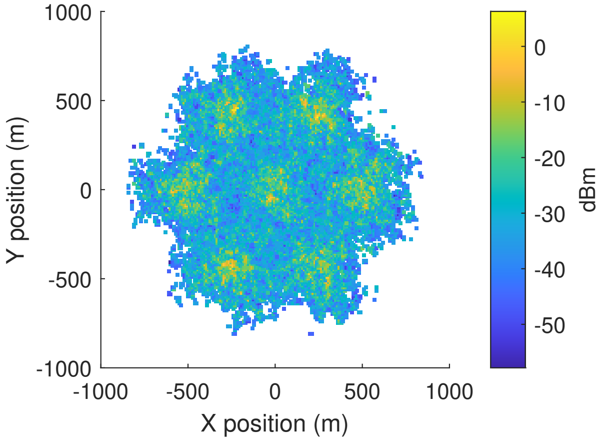


Fig. 5: MBSFN: Wide-band Signal

by the solid cyan lines, and the interference from the outer two rings, represented by the dashed red lines. The signals from the cells inside the MBSFN area can also lead to inter-symbol interference if it surpasses the extended CP, which is used for MBSFN in our simulations. A detailed analysis has been given in the work [8]. The varieties of network layouts and the signal combination from multiple locations make it too complicated for conventional performance analysis. Therefore, we seek simulations to provide insights into the performance in the designed scenarios.

III. Performance Analysis Over MBSFN Area and UE Locations

A. *Wide-band Signal and Interference Comparison*

Due to the multi-cell signal combination and interference reduction effect, the wide-band signal and interference can give us a more straightforward picture of the MBSFN behaviors. The signal power in dBm (decibels (dB) with reference to one milliwatt) for MBSFN and unicast are shown in Figure 4 and Figure 5, respectively. Overall, they share a very similar pattern with minor higher values for MBSFN, which is contributed by the signal combination from the 21 MBSFN cells.

Contrary to the signal power, the interference level differs significantly between unicast and MBSFN, as shown in Figure 6 and Figure 7. In unicast, the structure of the individual sites is recognizable, whereas in MBSFN the interference obliterates this pattern and leads to one common center. This is because the MBSFN cells are no longer generating interference to neighbor MBSFN cells, especially at cell edges.

In addition to the interference pattern's shift, the values of the interference level improve tremendously in MBSFN. For the total of 15 058 simulated points in Figure 6 and Figure 7, the mean of interference is reduced from -33.76 dBm in unicast to -39.98 dBm in MBSFN, with 95% confidence interval of [-33.88, -33.65] and [-40.09, -39.87] respectively; and the maximum of interference is reduced from 4.53 dBm to -11.84 dBm. These numbers imply that the interference reduction from neighboring cells in MBSFN could improve the UEs' SINR, especially at the center of the MBSFN area. In the majority of locations, this improvement outweighs the improvement from the signal combination discussed previously.

B. *Unicast Performance in the Selected Cells*

In unicast transmissions, we assume there is no cooperation or CoMP among cells. The performance of different cells should hence be similar. We present the results for this scenario in this section and use them as a benchmark for MBSFN comparisons in the next section. In this study, UEs are dropped at every pixel in the selected seven cells (Figure 1). Table I lists number of UEs dropped at each cell.

Since in multicast, only one single layer is supported in LTE, for fair comparisons, in unicast, we represent the signal portion by single-layer effective post-equalization (SLEPE)

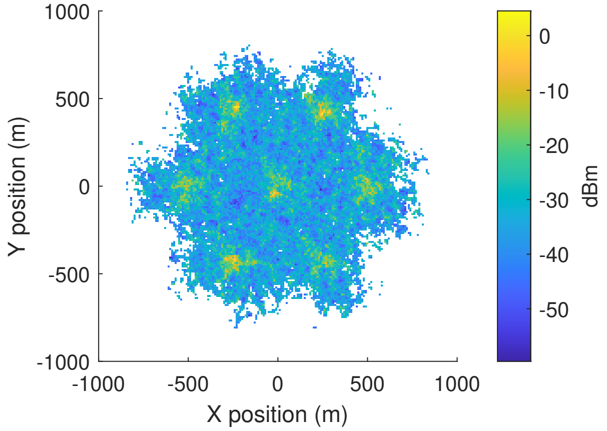


Fig. 6: Unicast: Wide-band Interference

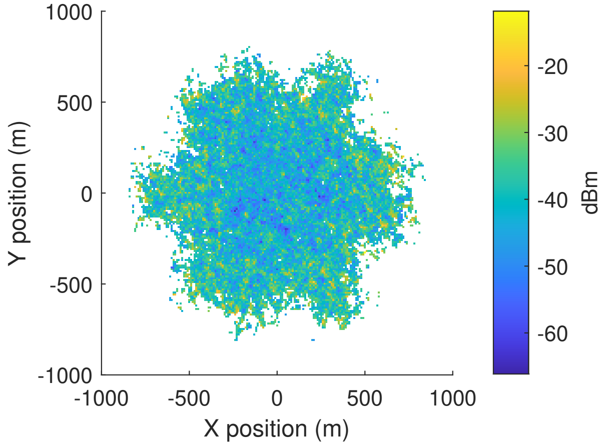


Fig. 7: MBSFN: Wide-band Interference

TABLE I: Cell Size by the Number of UEs

Cell	1	2	3	4	5	6	8
Number of UEs	683	697	820	700	660	735	770

SINR, which is calculated based on the Mutual Information Effective SINR Mapping (MIESM) method [12]. For each UE, we calculate the average SLEPE SINR over time by its arithmetic mean. The Cumulative Distribution Function (CDF) plot is shown in Figure 8. Combined with the number of UEs in each cell as in Table I, it can be seen that the distributions of the unicast UE SINR are quite similar among these cells. This result confirms that unicast cells with similar channel and pathloss models offer comparable performances. Figure 9 further captures the heatmap of the UE SLEPE SINR. It clearly shows the tri-sector site pattern, and that all cells share similar SINR distributions location-wise.

C. *MBSFN Performance in the Selected Cells*

In the case of MBSFN, the post-equalization SINR averaged over time for each UE in the seven selected cells are shown

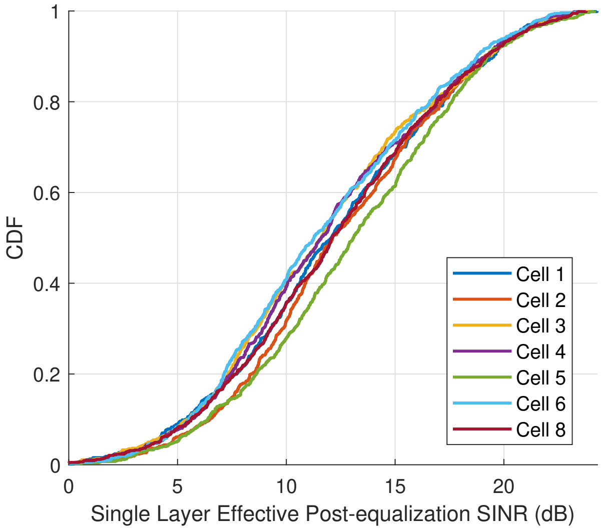


Fig. 8: Unicast: UE Single Layer Effective Post-equalization SINR (dB)

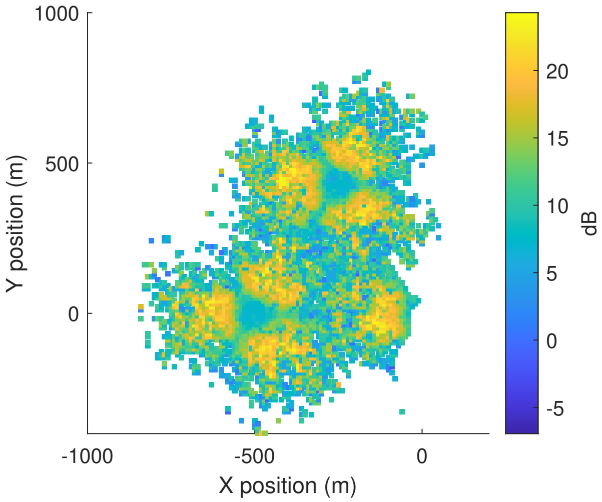


Fig. 9: Unicast: UE SINR Heatmap

in Figure 10. Compared with unicast in Figure 8, UEs in all of the selected cells have significant SINR improvements. For instance, the median SINR in unicast is around 12 dB, while the 25 percentile of SINR in MBSFN is already above 15 dB. This is due to the significant MBSFN gain discussed earlier. One thing to note is that, in reality, further improvement of SINR above a certain threshold has no contribution to the real performance. This is because MBSFN can only use a single data stream, and the maximum MCS caps the required SINR. This threshold is approximately 25 dB in our simulation [13]. For the same reason, only SINR below 30 dB is shown in Figure 10 for MBSFN.

Also, it can be observed that there are more performance

Shen, Chen; Liu, Chunmei; Rouil, Richard A.; Ah-Choi, Hyeong. "Study of Multicast Broadcast Single Frequency Network Area in Multicast Communication." Presented at 14th International Conference on Signal Processing and Communication Systems (ICSPCS 2020). December 14, 2020 - December 16, 2020.

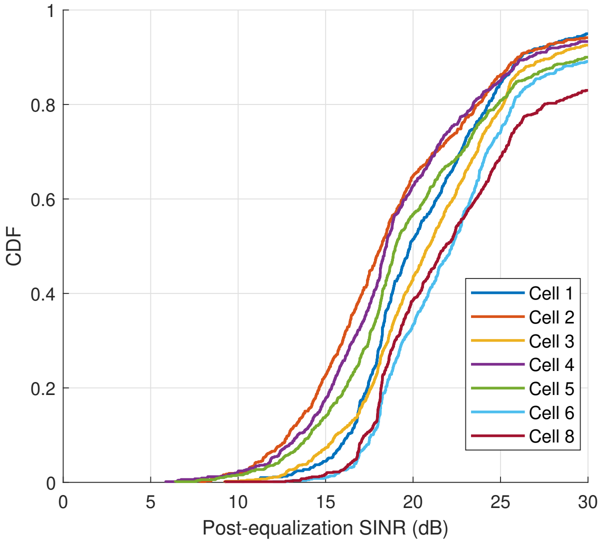


Fig. 10: MBSFN: UE Post-equalization SINR (dB)

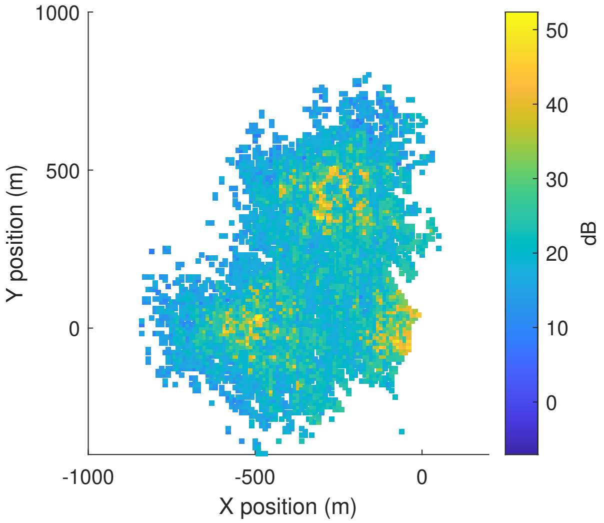


Fig. 11: MBSFN: UE SINR Heatmap

variations among cells in MBSFN than in unicast. Some cells have a constant performance advantage over other cells. For example, the SINR in Cell 8 has approximately 4 dB improvement over Cell 2 and Cell 4. From Figure 2, Cell 8 is at the center of the MBSFN area, while Cell 2 and 4 are at the edge. That is, the MBSFN gain decreases from the MBSFN center to the edge. This conclusion can be further verified by the other cells. For example, if we sort the seven cells by how close it is to the center of the MBSFN area visually, we should get the order: 8, 6, 3, 1, 5, 4, and 2 (Cell Index). This order is consistent with the order of SINR CDF curves in Figure 10. This is because the MBSFN area center is further away from interference sources outside the MBSFN area (the outer rings), and at the same time, receives stronger constructive signals within the MBSFN area.

Also, the tri-sector pattern that exists in unicast is weakened in the case of MBSFN because of the different signal and interference distributions, as shown in subsection III-A, which draws an essential difference from the unicast.

D. Other MBSFN Deployment Scenarios

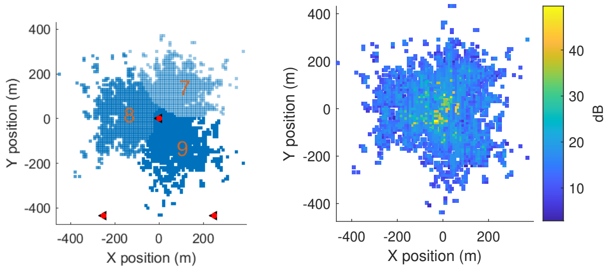
To further explore the impact of the MBSFN area, we analyze different MBSFN area sizes and deployment shapes in terms of the post-equalization SINR. Instead of the 21-cell MBSFN configuration used above, the MBSFN areas in this sub-section are as follows.

First, we introduce two cases with 3-cell MBSFN, whose area consists of 3 MBSFN cells but in different deployment shapes. Case 1 shown in Figure 12a has 3 cells of one site for the MBSFN area, whereas case 2 shown in Figure 14a has 3 cells from 3 different sites that make up the MBSFN area. Both cases share the same area shape and number of cells but differ in cell radiation directional pattern. If we set a cell with a radiation direction, case 1 is centrifugal, whereas case 2 is similar to centripetal.

The post-equalization SINR in these 2 cases have very distinctive distributions, as shown in the SINR heatmaps in Figure 12b and Figure 14b, as well as the SINR CDFs in Figure 13 and Figure 15. We can observe that case 1 SINR has a concentric hierarchy pattern, while it is more evenly distributed in case 2. The maximum SINR in case 1 can reach 45 dB, accompanied by a significant SINR portion below 15 dB. In contrast, the majority SINR in case 2 is between 15 dB and 25 dB.

This significantly different SINR distribution tells us that, even with the same public safety incident area, its location relative to the network could lead to significant performance differences. Recall that MCS used for MBSFN transmissions is selected to be the minimum MCS over all UEs. This result can provide very insightful guides to deploy mobile base stations when public safety incidents are out-of-coverage of macro-cells. For instance, with the same incident area size and shape, if first responders are concentrated within a small area, deploying mobile base stations as in case 1 would lead to much higher MBSFN MCS applied. In contrast, if first responders are evenly located across a relatively large area, mobile base station deployments based on case 2 is preferred to achieve higher MBSFN MCS since minimum MCS of all UEs can achieve relatively good value.

The next set of cases consists of three sites or nine cells as the MBSFN area. In case 3, three sites are deployed along a line, as shown in Figure 16a, and in case 4, three sites are arranged to form a triangle, as shown in Figure 18a. The SINR heatmaps in Figure 16b and Figure 18b, together with their CDFs in Figure 17 and Figure 19, show that the cells in case 3 share similar SINR distribution, whereas the cells in case 4 are more dispersed with Cell 1, 6, and 8 having better SINR over other cells. This is reasonable since they are located closer to the ‘center’ of the MBSFN area and experiencing higher



(a) Cell Map

(b) SINR Heatmap

Fig. 12: MBSFN 3-cell Map, Case 1

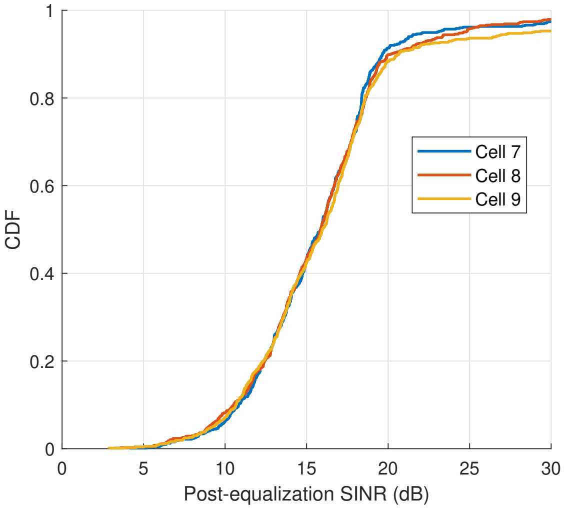
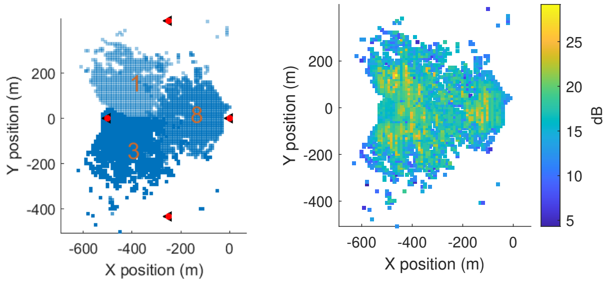


Fig. 13: MBSFN 3-cell CDF, Case 1

MBSFN gain, especially the portion coming from interference reduction. Unlike the case set with three cells where both cases are advantageous when serving some specific conditions, in this scenario, case 4 outperforms case 3 comprehensively.

This result tells us that even if public safety incidents cover the same area size, their different area shapes will lead to different performances. Note that by comparing SINR CDFs with MBSFN area size of 3, 9, or 21 cells shown earlier, we can see that different MBSFN area sizes lead to a different performance, too, with more considerable SINR improvement with larger area size.



(a) Cell Map

(b) SINR Heatmap

Fig. 14: MBSFN 3-cell Map, Case 2

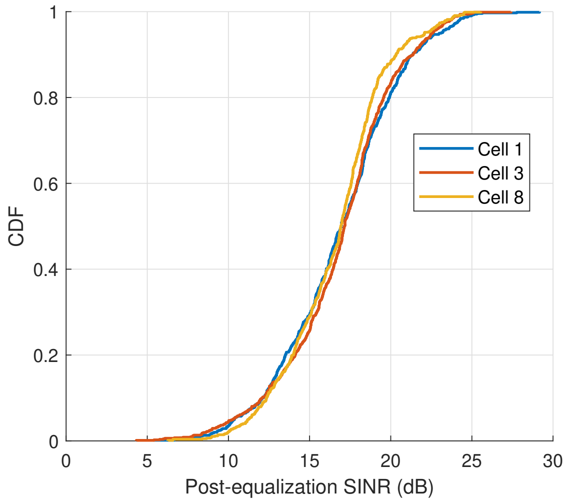
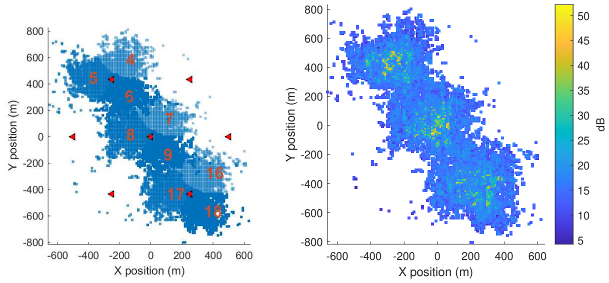


Fig. 15: MBSFN 3-cell CDF, Case 2



(a) Cell Map

(b) SINR Heatmap

Fig. 16: MBSFN 3-site Map, Case 3

To sum up, the simulation results in this section can be directly used to predict and evaluate performance when there exists an underlying macro-cell network in the public safety incident area. When the incident is out-of-coverage of macro-cell networks, the results can be used to guide mobile base station deployments.

IV. Conclusion

In this paper, we have investigated UE post-equalization SINRs across MBSFN area with different area sizes and shapes. With the introduction of the multicast mechanism, we first showed the different patterns of wide-band signal and interference in MBSFN versus unicast, where the wide-band interference of the 21-cell MBSFN area forms a concentric scheme with the lowest interference at the center. We then compared the selected cells from unicast and 21-cell MBSFN, where the post-equalization SINR improves dramatically in MBSFN while showing distinct distributions among cells,

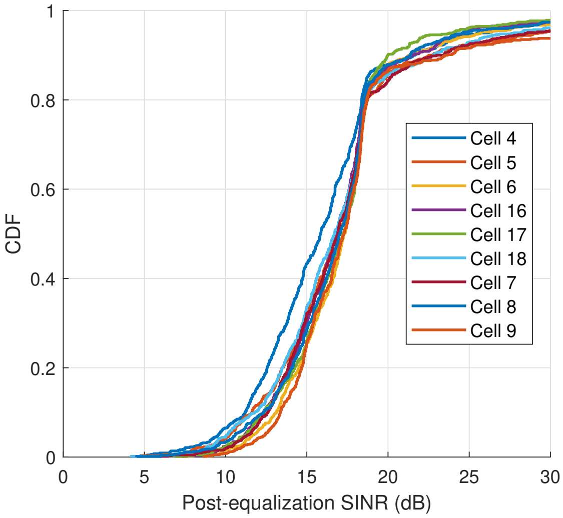
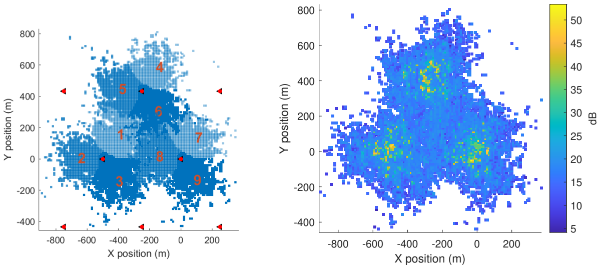


Fig. 17: MBSFN 3-site CDF, Case 3



(a) Cell Map

(b) SINR Heatmap

Fig. 18: MBSFN 3-site Map, Case 4

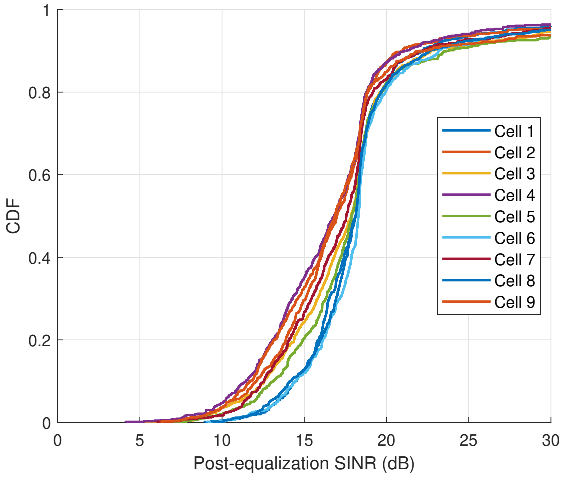


Fig. 19: MBSFN 3-site CDF, Case 4

which further confirms the previous wide-band observations. Moreover, we explored different deployment locations and area shapes with the same number of MBSFN cells, 3-cells and 9-cells over 3-sites, respectively. We showed that not only MBSFN area size impacts performance, but also MBSFN area shape and its relative locations to the network. The results obtained can provide insightful assessments for real case MBSFN performance and guides for mobile base station deployments in case of out-of-coverage.

References

[1] 3GPP TS23.246, "Multimedia Broadcast/Multicast Service (MBMS) Architecture and Functional Description," 3GPP, Standard, 9 2019.

[2] 3GPP TS25.346, " Technical Specification group radio access network; introduction of the multimedia broadcast multicast service (MBMS) in the radio access network (RAN); stage 2 (release 9)," 3GPP, Standard, 9 2009.

[3] K. Kirev and S. Schwarz, "Investigation of optimal mcs and subcarrier spacing in mbsfn systems," in *WSA 2020; 24th International ITG Workshop on Smart Antennas*, 2020, pp. 1–6.

[4] L. Zhang, Y. Cai, Z. He, C. Wang, and P. Skov, "Performance evaluation of lte mbms baseline," in *2009 5th International Conference on Wireless Communications, Networking and Mobile Computing*, 2009, pp. 1–4.

[5] 3GPP TS36.101, "Evolved Universal Terrestrial Radio Access (E-UTRA; user equipment (ue) radio transmission and reception," 3GPP, Standard, 6 2020.

[6] A. Alexiou, C. Bouras, V. Kokkinos, A. Papazois, and G. Tsichritzis, "Efficient mcs selection for mbsfn transmissions over lte networks," in *2010 IFIP Wireless Days*, 2010, pp. 1–5.

[7] 3GPP TS36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures," 3GPP, Standard, Jan. 2019.

[8] C. Liu, C. Shen, J. Chuang, R. A. Rouil, and H. Choi, "Evaluating unicast and mbsfn in public safety networks," in *IEEE PIMRC 2020 - 2020 IEEE International Conference on Communications*, Aug. 2020.

[9] C. Borgiattino, C. Casetti, C. . Chiasserini, and F. Malandrino, "Efficient area formation for lte broadcasting," in *2015 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2015, pp. 202–210.

[10] C. Shen, M. Yun, A. Arora, and H.-A. Choi, *Efficient Mobile Base Station Placement for First Responders in Public Safety Networks*, 01 2020, pp. 634–644.

[11] C. Shen, M. Yun, A. Arora, and H. Choi, *Dynamic Placement Algorithm for Multiple Classes of Mobile Base Stations in Public Safety Networks*, 08 2019, pp. 112–125.

[12] Z. Hanzaz and H. D. Schotten, "Analysis of effective sinr mapping models for mimo ofdm in lte system," in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2013, pp. 1509–1515.

[13] 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.

International Workshop on Deep Video Understanding

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

This is the introduction paper to the International Workshop on Deep Video Understanding, organized at the 22nd ACM Interational Conference on Multimodal Interaction. In recent years, a growing trend towards working on understanding videos (in particular movies) to a deeper level started to motivate researchers working in multimedia and computer vision to present new approaches and datasets to tackle this problem. This is a challenging research area which aims to develop a deep understanding of the relations which exist between different individuals and entities in movies using all available modalities such as video, audio, text and metadata. The aim of this workshop is to foster innovative research in this new direction and to provide benchmarking evaluations to advance technologies in the deep video understanding community.

CCS CONCEPTS

• **Information systems** → *Multimedia content creation*; **Multimedia and multimodal retrieval**.

KEYWORDS

video understanding; multimedia; multimodal interaction; information retrieval; video ontology

ACM Reference Format:
Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. International Workshop on Deep Video Understanding. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3382507.3419746

1 WORKSHOP INTRODUCTION

Deep video understanding is a difficult task which requires systems to develop an insightful analysis and understanding of the relationships among different entities in video, to use known information to reason about other, more hidden information, and to populate a knowledge graph (KG) with all acquired information. There has been efforts to encourage research in high level video understanding such as the "MovieQA" [6] and "The Large Scale Movie Description Challenge" [5]. 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 [4] 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. A large-scale dataset of corresponding movie trailers, plots, posters, and metadata was developed by [1] 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.

To work on this task, a system should take into consideration all available modalities (speech, image/video, and in some cases text). The aim of this workshop is to push the limits of multimodal extraction, fusion, and analysis techniques to address the problem of analyzing long duration videos holistically and extracting useful knowledge to utilize it in solving different types of queries. The target knowledge includes both visual and non-visual elements. As videos and multimedia data are getting more and more popular and usable by users in different domains, the research, approaches and techniques we aim to be applied in this workshop will be very relevant in the coming years and near future. The call for contributions in this workshop supported long, short and abstract papers related to multimedia understanding, in addition to an optional track for researchers who are interested to apply their techniques on a new pilot creative commons movie dataset (HLVU)[3] collected by the NIST, who distributed the dataset with development data, testing queries, and finally evaluated and scored the submitted runs by participating researchers. In this optional challenge track, all participants were invited to submit a paper describing their approaches to solve the testing queries. The following sections summarize the two tracks of the workshop.

1.1 Track 1

In this track authors were invited to apply their approaches and methods on a novel High-Level Video Understanding (HLVU) dataset made available by the workshop organizers which included 10 movies with a Creative Commons (CC) license. These movies were annotated by human assessors and full ground truth, including the Knowledge Graph of all entities and relationships, was made available for 6 of these movies which made up the development set.

[*] Georgetown University

Curtis, Keith; Awad, George; Rajput, Shahzad; Soboroff, Ian. "International Workshop on Deep Video Understanding." Presented at 22nd International Conference on Multimodal Interaction, Utrecht, NL. October 25, 2020 - October 29, 2020.

Evaluation and scoring was supported for two main query types distributed with the dataset:

(1) Multiple choice question answering on Knowledge Graph for selected movies.
(2) Path analysis between persons / entities of interest in a Knowledge Graph extracted from selected movies.

Movies were annotated by human assessors who drew a Knowledge Graph of nodes. Nodes were comprised of important persons, entities, and major concepts from the storyline. Edges connecting these nodes were of the relationship type between persons, entities, and concepts. For example if *Person A* was a Student at *University*, there was an edge named *Studies At* connecting *Person A* to *University*. Annotators were given a predefined ontology of most common relationships between people (family-based, workplace-based, social, etc). Graphical examples of this are available at [3] and is available online at [2].

Using this annotation scheme, a set of multiple choice questions were generated for the test set movies on the relationships between various nodes in this KG. In addition, the full path connections between various nodes were analysed and participants were asked to submit possible paths. For example, a path question would ask how is person X connected to person Y and systems were required to return back all of the possible paths with correct relationships between the two persons. Further details on this annotation scheme, question types, metrics, and movies used for this dataset are available from [3].

1.2 Track 2

In this track authors were invited to submit contributions related, but not limited, to the following topics applied on the provided HLVU dataset or any external datasets:

- Multimodal feature extraction for movies and extended video
- Multimodal fusion of computer vision, text/language processing and audio for extended video / movie analysis
- Machine Learning methods for movie-based multimodal interaction
- Sentiment analysis and multimodal dialogue modeling for movies
- Knowledge Graph generation, analysis, and extraction for movies and extended videos

2 WORKSHOP CONTENT

2.1 Keynote Speakers

In addition to authors of accepted papers, we have invited three keynote speakers bringing different perspectives to this challenging research direction:

- **Klaus Schoeffmann (Klagenfurt University, Austria)**
  Dr. Klaus Schoeffmann is an Associate Professor at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. His research focuses on video content understanding (in particular of medical/surgery videos), multimedia retrieval, interactive multimedia, and applied deep learning.
- **Dima Damen (University of Bristol, UK)**
  Dr. Dima Damen is a Reader (Associate Professor) in Computer Vision at the University of Bristol, United Kingdom. Her research interests are in the automatic understanding of object interactions, actions and activities using static and wearable visual (and depth) sensors.
- **Koichi Shinoda (Tokyo Institute of Technology, Japan)**
  Prof. Koichi Shinoda is a professor with the Tokyo Institute of Technology, Japan. His research interests include speech recognition, video information retrieval, statistical pattern recognition, and human interfaces.

2.2 Papers

At the time of writing this introduction paper, the workshop paper submissions were still open. We had one participant in Track 1 of this workshop, and a number of interesting works submitted for Track 2 of this workshop. We hope that this will provide a good balance of interesting work related to this research area. However, due to the fact that paper reviews had yet to be completed we are unable to provide further information on accepted papers at this stage.

3 WORKSHOP ORGANIZATION

3.1 Organizing Committee

- Keith Curtis (National Institute of Standards and Technology, USA)
- George Awad (Georgetown University & National Institute of Standards and Technology, USA)
- Shahzad Rajput (Georgetown University & National Institute of Standards and Technology, USA)
- Ian Soboroff (National Institute of Standards and Technology, USA)

4 CONCLUSIONS

There is a good diversity of research presented at this workshop, and we are satisfied that we have encouraged further work in this new direction. The research presented at this workshop coupled with analysis of what has been learned during the annotation of this dataset and running the benchmark campaign has provide a good basis for the extension and continuation of this research into the next few years.

ACKNOWLEDGMENTS

The authors would like to thank the Information Technology Laboratory at NIST for sponsoring all annotation work done on the HLVU dataset and sponsoring the first three organizers of this workshop. Also we thank the ICMI workshop chairs and main conference for including this new workshop, and all paper authors for their research efforts and contributions.

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] 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).

[2] Keith Curtis and George Awad. 2020 (accessed August 26, 2020). *DVU Challenge*. https://drive.google.com/drive/folders/1q1Ca0aFJrF9tB8hsw-mrI9d4tzy5wlPZ

[3] 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*. 355–361.

[4] 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.

[5] 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.

[6] 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.

SiF 2020– The 11th International Conference on Structures in Fire
The University of Queensland, Brisbane, Australia, November 30 – December 2, 2020

THERMAL RESPONSE AND CAPACITY OF BEAM END SHEAR CONNECTIONS DURING A LARGE COMPARTMENT FIRE EXPERIMENT

Xu Dai[1,*], Lisa Choe[2], Erica Fischer[3], Charles Clifton[4]

ABSTRACT

The role of steel connections is essential in structural fire design and analysis for steel-framed composite structures. The current structural design provisions provide strength reduction factors of load-carrying members and their end-connection elements (e.g. bolts) at elevated temperatures, based on small-scale experiments under uniform heating conditions. The realistic temperature evolution in member connections, especially as part of full-scale floor assemblies exposed to a large compartment fire, has not been well characterized. A large compartment fire experiment was recently conducted on a 9.1 m by 6.1 m composite floor assembly as part of a two-story steel framed building. The test assembly had a total of ten shear-tab (fin-plate) connections subjected to combined fire and mechanical loading. This paper presents the measured thermal response of these connections in comparison with the corresponding Eurocode 3 predictions with two methods (1) incorporating the beam bottom flange temperature at midspan and (2) the section factor method. The results show that the Eurocode 3 methods conservatively predict the maximum temperature during heating and the cooling rate but overestimate the high-temperature strength of connections while using the section factor method. The predicted thermal responses are highly influenced by the fire protection sprayed on the connection region which was actually at least 43% thicker than the protection on the beams used in this test program. However, partial shear failure of bolts was witnessed in the test. This suggests that designing connections solely through temperature provisions may not guarantee a safe structural fire design. The axial load demand of the shear connection due to restraints to thermal elongation or contraction should be considered in future design guidance.

Keywords: Connections; fire test; Eurocodes; temperatures; capacity; shear tab

1 INTRODUCTION

In fire safety design of buildings, the structure is required to meet the desired performance objectives. At a minimum these include maintaining structural integrity such that compartmentalization is not compromised [1]. However, the accurate quantification of fire severity in a structure has been challenging due to lack of design tools validated against experimental data. Particularly, the thermal response of composite beam connections to a large compartment fire is one of such examples [2-4]. Currently in Eurocode 3 [5] and other similar provisions used in many other countries including the United States, the thermal gradient of the beam end connections is estimated using the empirical equations based upon the bottom flange temperature of composite floor beams. However, the bottom flange beam temperature utilized in these

[1] Foreign Guest Researcher, National Institute of Standards and Technology (NIST)
e-mail: xu.dai@nist.gov, *corresponding author, ORCID: https://orcid.org/0000-0002-9617-7681

[2] Research Structural Engineer, National Institute of Standards and Technology (NIST)
e-mail: lisa.choe@nist.gov, ORCID: https://orcid.org/0000-0003-1951-2746

[3] Assistant Professor, Oregon State University
e-mail: erica.fischer@oregonstate.edu, ORCID: https://orcid.org/0000-0002-7653-2068

[4] Associate Professor, University of Auckland
e-mail: c.clifton@auckland.ac.nz, ORCID: https://orcid.org/0000-0003-0723-1699

Dai, Xu; Choe, Lisa; Fischer, Erica; Clifton, Charles. "Thermal response and capacity of beam end shear connections during a large compartment fire experiment." Presented at 11th International Conference on Structures in Fire (SiF 2020), Brisbane, AU. November 30, 2020 - December 02, 2020.

methods are assumed to be remote from the connection. The thermal gradient across the beam-end connections is then used to calculate the connection capacity incorporating strength reduction factors for bolts and welds. The reliability of those Eurocode 3 predictions is unknown especially for the fire-protected connections and full-scale typical beam sizes and spans within a large compartment fire [6, 7] due to lack of such experimental temperature data. Furthermore, even if the thermal response of the steel connection components was within the range predicted by Eurocode 3 methods, to what extent this will ensure a robust structural fire design is also uncertain, based on the lack of supporting experimental data and analysis.

The work presented in this paper aims to: (1) enrich the experimental data library for thermal and structural response of the shear-tab (fin-plate) connections with fire protection subjected to a full-scale large compartment fire; (2) examine the Eurocode 3 connection design (temperature and strength) utilizing experimental data [8] with a large compartment fire; and (3) identify gaps in knowledge or data for structural design of shear connections under fire conditions.

2 FIRE TEST

2.1 The test building and fire compartment

A two-story steel frame with composite floors has three by two bays in plan (18 m × 11 m) with a total building height of 7.2 m, and the fire was in a large compartment at the south-central bay having dimensions of 9.1 m × 6.1 m with a 3.8 m ceiling height (Figure 1). The composite floor assemblies were designed to resist an ambient design gravity load of 8.6 kPa. The test floor assembly was subjected to a floor load of 5.3 kPa following the ASCE 7 [9] load combination for extraordinary events (1.2 × dead load + 0.5 × live load).

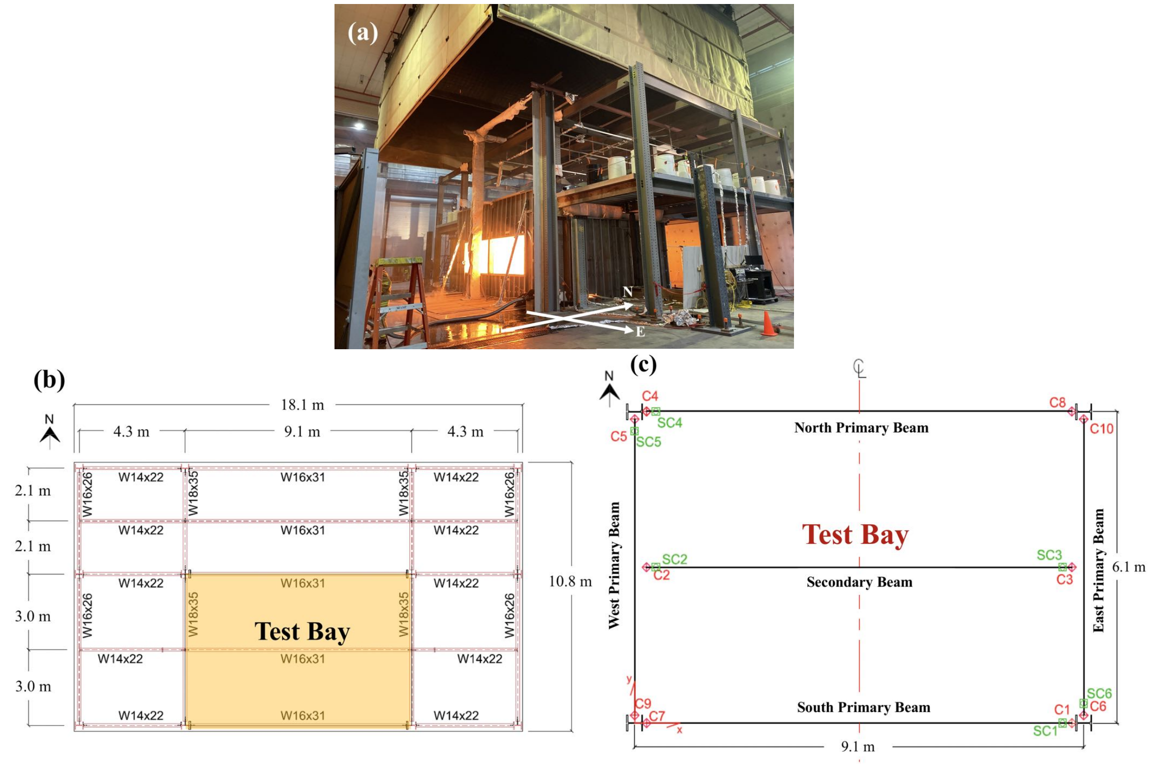


Figure 1. NIST large compartment fire experiment [8] on a steel-composite building. (a). Photo taken during the experiment; (b). Plan view of the test building and test bay location; and (c). Plan view of composite connection instrumentation locations in the experimental compartment, ten in total, from connection 1-10 (abbrev. as C1-C10).

For passive fire protection of exposed steel members, a medium density (ranging from 240 kg/m^3 to 350 kg/m^3 [10]) gypsum-based cementitious material, was sprayed on beams and connections exposed to fire. The design thickness of insulation on both south and north primary W16×31 beams, as well as on the west and east primary W18×35 beams inside the fire test compartment was 17.5 mm (11/16 inch) determined using the Underwriter Laboratory (UL) directory N791 [11] for the 2-hour restrained beam rating. Insulation thickness of the secondary W16×31 beam was 13 mm (7/16 inch), slightly thinner than the primary beams, determined using UL D949 [12] for the 2-hour restrained assembly rating. The exposed connection regions and columns were over sprayed with the same insulation material with the thickness of 25 mm or greater, i.e. 43% thicker than the primary beams, to ensure the 3-hour rating of columns.

2.2 The connections and relevant instrumentations

Two types of simple shear connections were used in this test program: standard shear tabs for the beam-to-column flange and beam-to-beam web connections; extended shear tabs for the beam-to-column web connections. All shear tabs were 9.5 mm in thickness and made of ASTM A36 [13] steel (the minimum yield stress of 245 MPa). The size of fillet welds was 6.3 mm. All structural bolts (Gr. A325 specified in the ASTM F3125 [14]) had a diameter of 19 mm. The dimensions of the short-slot holes (21 mm in width and 25 mm in length) drilled on the shear tabs conform to the AISC 360 specification [1]. Examples of the connections and mounted thermocouples[5] are demonstrated in Figures 2(a) and 2(b).

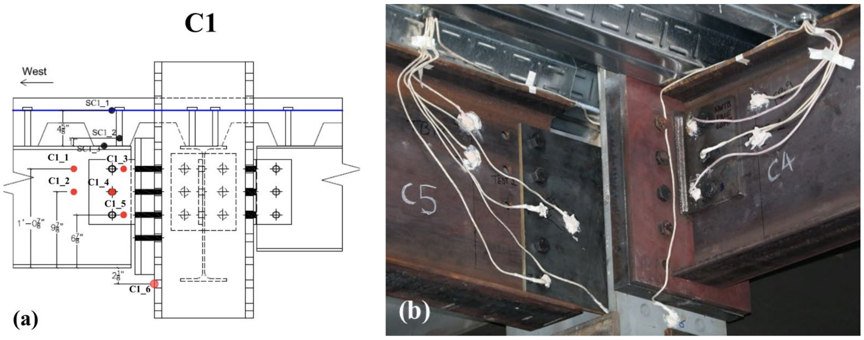


Figure 2. (a). Thermocouple instrumentation locations at connection C1, tagged as C1_1 to C1_6; and (b). Thermocouple instrumentation locations at connection C4 (standard shear tab) and connection C5 (extended shear tab), prior to SFRM installation.

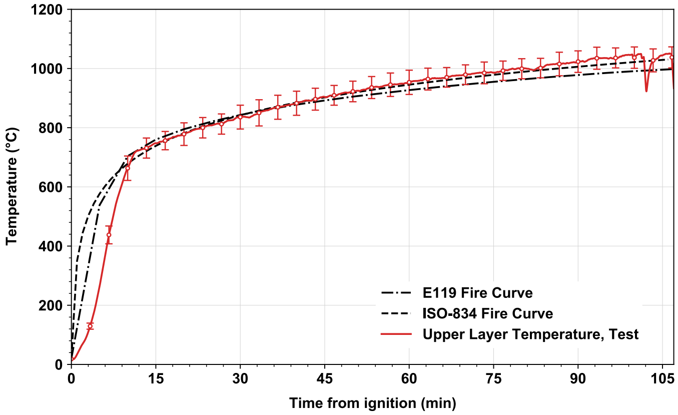


Figure 3. A comparison between measured[6] upper layer gas temperature within the test compartment, and standard fire curves (i.e. the ASTM E119 fire curve, and the ISO-834 fire curve)

[5] An expanded uncertainty of thermocouple locations is estimated to be ± 6 mm with a coverage factor of 2 (95% confidence interval)

[6] An expanded uncertainty of measured gas temperatures is estimated to be ± 8% at 1100 °C with a coverage factor of 2

2.3 Mechanical and fire loading

The vertical shear load imposed on the connections was in the range of 0.2 to 0.4 of their ambient design capacities during the fire experiment. The fire load (or exposure) was applied using natural gas burners [15-17] following the ASTM E119 temperature-time curve [18] lasting 107 mins equivalent to 921 MJ/m^2 with ± 1.5 MJ/m^2 uncertainty (95 % confidence interval) as the applied fire load density. At 107 min the heat release rate reached to its maximum value of 10.8 MW. As demonstrated in Figure 3, the test revealed that the measured time-temperature curve matched better with the ISO Standard Fire curve [19] (which is more severe than the ASTM E119 curve) after 45 mins. The maximum standard deviation of the upper layer gas temperature was around 70 °C throughout the heating regime. This standard fire environment was to ensure that the whole structural assembly was being challenged to a severe fire impact and for the research interest of steel connections. This impact included a large compressive force induced by the restraint to thermal expansion during the heating phase, a tensile force due to catenary action during heating, and a tensile force by the restraint to thermal contraction during cooling.

2.4 Gas temperature and steel temperatures

An example of the temperature comparison is presented in Figure 4, including upper layer gas temperature and temperatures of the west primary beam bottom flange (average of thermocouple TBi_5 and TBi_6) and connection C5. This comparison considers seven hours of testing duration including the natural cooling phase. After the burner was switched off at 107 min, the upper layer gas temperature decreased sharply from 1040 °C to 490 °C within approximately 10 mins into cooling. This drastic temperature decrease would likely induce large tension forces on the connections, due to the restraint to thermal contraction of the steel beam during the cooling stage.

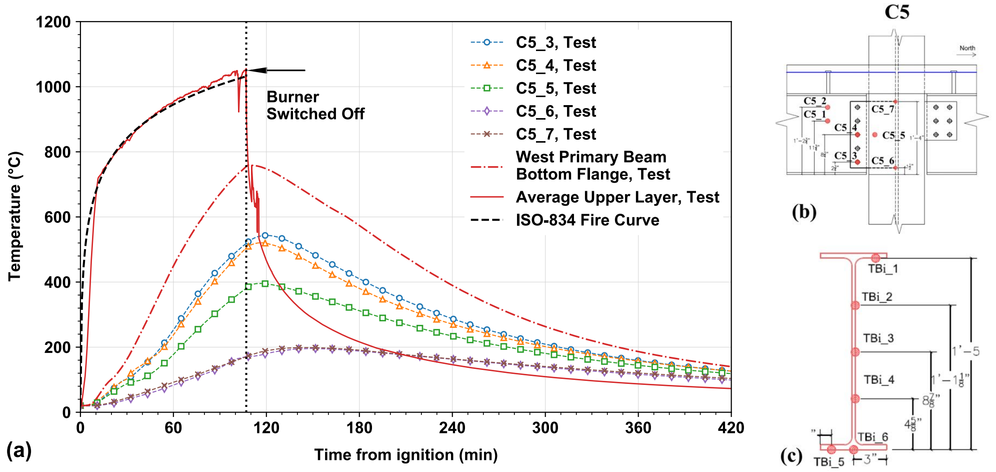


Figure 4. (a). Comparison of the time temperature curves for the compartment upper layer gas, west primary beam bottom flange, and steel connection[7] C5; (b). Thermocouple instrumentation locations at connection C5, tagged as C5_1 to C5_7; and (c). Thermocouple instrumentation locations at west primary beam midspan, tagged as TBi_1 to TBi_6.

The bottom flange temperature of the west primary beam reached its measured maximum value of 760 °C and cooled off gradually due to the presence of the fire protection. The connection temperatures followed a similar tendency but with relatively lower maximum values (e.g. 540 °C at C5_3) and a time delay to their peaks, provided that thicker SFRM was applied. The test showed that C5_3 and C5_4 (bolts) indicated higher component temperatures, compared to the measured temperatures at C5_6 and C5_7 (welds). This is believed to be because C5_6 and C5_7 were affected by the thermal shadow effect or the heat conduction loss to the connected column at this region. Those relative relationships of temperatures were further evaluated using Eurocode 3 methods, as detailed in the subsequent section.

[7] An expanded uncertainty of measured steel temperatures is estimated to be ± 4% at 970 °C with a coverage factor of 2

3 CONNECTION TEMPERATURES

3.1 Eurocode 3 method

According to the Eurocode 3, for beam-to-beam and beam-to-column connections where concrete floors are atop the beams, temperatures of the connections can be estimated based upon the bottom flange temperature of the connected steel beam at midspan. Considering the depths of the steel members used in the experiment (i.e. W16×31 and W18×35), are both greater than 400 mm, hence two equations are used:

when h is less or equal than $D/2$:

$$\theta_h = 0.88\theta_o \qquad (1)$$

when h is greater than $D/2$:

$$\theta_h = 0.88\theta_o\,[1 + 0.2(1\text{-}2h/D)] \qquad (2)$$

where θ_h is the temperature at height h (mm) of the steel beam, see Figure 5;

θ_o is the bottom flange temperature of the steel beam remote from the connection;

h is the height of the component being considered above the bottom of the beam in (mm);

D is the depth of the beam in (mm).

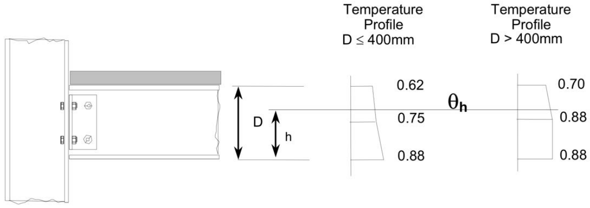


Figure 5. Thermal gradient within the depth of a composite connection (figure adapted from Eurocode 3 [5]).

If the gas temperature is known, the Eurocode 3 step-by-step section factor method can be used to estimate the steel beam bottom flange temperature θ_o as follows:

$$\Delta\theta_{a,t}=\frac{\lambda_p\, A_p/V(\theta_{g,t} - \theta_{a,t})}{d_p c_a \rho_a\,(1 + Ø/3)}\,\Delta_t - (e^{Ø/10}-1)\Delta\theta_{g,t} \qquad (3)$$

$$Ø=\frac{c_p\rho_p}{c_a\rho_a}\,d_p\,A_p/V \qquad (4)$$

where A_p/V is the section factor for steel members insulated by fire protection material in (m^{-1});

c_a is the temperature dependant specific heat of steel in (J/kgK);

c_p is the temperature independent specific heat of the fire protection material in (J/kgK);

d_p is the thickness of the fire protection material in (m);

λ_p is the thermal conductivity of the fire protection in (W/mK);

ρ_a is the unit mass of steel in (kg/m^3);

ρ_p is the unit mass of the fire protection in (kg/m^3);

$\theta_{a,t}$ is the steel temperature at time t in (^{o}C);

$\theta_{g,t}$ is the ambient gas temperature at time t in (^{o}C);

Δ_t is the time interval in (seconds);

As detailed above, theoretically two analytical methods can be used to estimate temperatures of connection components: if the steel beam bottom flange temperature at midspan is known, then equations (1) and (2) can be employed directly; if the gas temperature within the compartment is known, then equation (3) can be used to estimate the steel beam temperature which is a main variable of equations (1) and (2). The following section will examine those two methods, utilizing the experimental data to evaluate the applicability of the Eurocode 3 provisions.

3.2 Comparison between the measurements and Eurocode 3 predictions - method 1

Following the Eurocode 3 convention, see Figure 5, dimensionless thermal gradients of all the test connections are summarized in Figure 6. These gradients were estimated when the connected structural beam members reached to their maximum deflections, at around 107 min. At this time, as summarized in Table 1, temperatures of the end connections of the secondary and primary W16×31 beams were in excess of 500 °C, whereas the end connections of the primary W18×35 beams were heated below 400 °C. The temperature discrepancy between the primary and secondary beam-end connections is mainly due to the difference in applied SFRM thickness, see Table 1. As shown in Figure 5, for components of the standard shear tabs connecting W16×31 beams, the dimensionless experimental temperatures vary from 0.35 to 0.7, less than the values calculated using the Eurocode 3 (0.7 to 0.88). For those of the extended shear tabs at the ends of W18×35 beams, moreover, the measured dimensionless temperatures range from 0.5 to 0.86, which remain below 0.88 calculated using equations (1) and (2). This comparison demonstrates that the Eurocode 3 provisions, estimating temperatures of the standard shear tabs as a function of the beam bottom flange temperatures at midspan, are conservative. It is also anticipated that under a natural fire (rather than in a standard fire), the temperature difference between the connections and the beam flange at midspan is greater. If a natural fire is fuelled by array of wood cribs, e.g. [3], the upper layer gas temperature is expected to be less uniform and highly influenced by other factors (e.g. distribution of fuel, ventilation, wood properties) when compared with the conditions in which the test fire is controlled by natural gas burners.

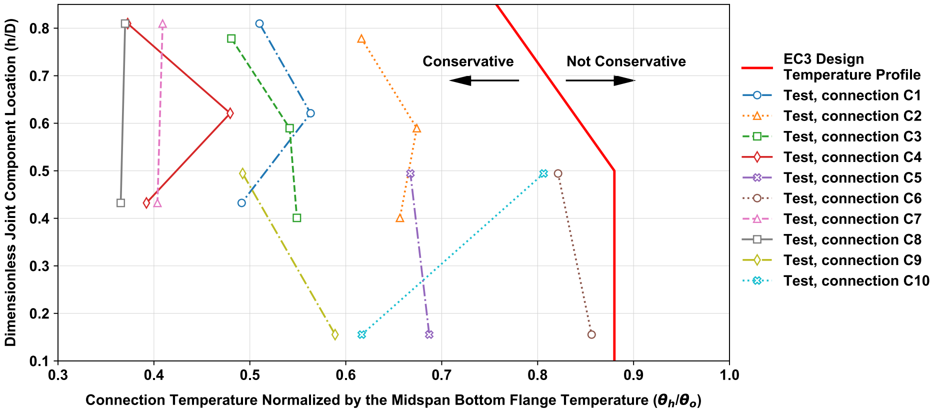


Figure 6. Thermal gradient within the depth of the composite connections, test vs. Eurocode 3

Table 1. Summary of the connections: connected members, types, SFRM thickness (± represented as standard deviation), and measured max component temperature.

Connection Tag	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**
Connected Members	primary beam - column flange	secondary beam - primary beam	secondary beam - primary beam	primary beam - column flange	primary beam - column web	primary beam - column web	primary beam - column flange	primary beam - column flange	primary beam - column web	primary beam - column web
Member Dimension	W16×31	W16×31	W16×31	W16×31	W18×35	W18×35	W16×31	W16×31	W18×35	W18×35
Shear Tab Type	Standard	Standard	Standard	Standard	Extended	Extended	Standard	Standard	Extended	Extended
Fire Protection Thickness (mm)	28 ± 2	31 ± 3	29 ± 3	26 ± 2	27 ± 3	25 ± 1	29 ± 6	28 ± 4	24 ± 3	25 ± 2
Maximum Temperature (°C)	400	660	570	380	530	590	300	310	500	580

Dai, Xu; Choe, Lisa; Fischer, Erica; Clifton, Charles. "Thermal response and capacity of beam end shear connections during a large compartment fire experiment." Presented at 11th International Conference on Structures in Fire (SiF 2020), Brisbane, AU. November 30, 2020 - December 02, 2020.

3.3 Comparison between the measurements and Eurocode 3 predictions - method 2

An example of comparison between the measured and predicted temperatures based upon the step-by-step section factor method (method 2) is presented in Figure 7. In this example, the connection components C1_4 and C4_4 were situated at the diagonal locations of the test compartment.

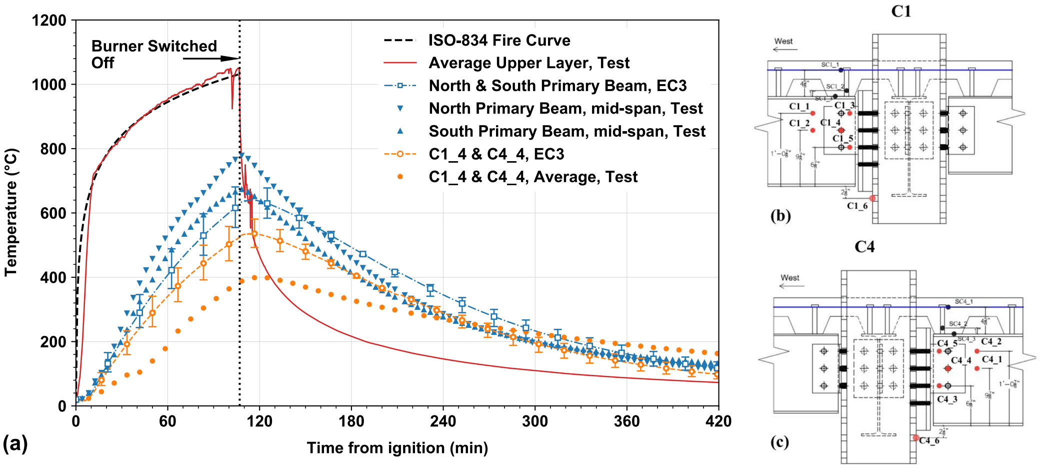


Figure 7. (a). Comparison between predicted temperatures with the measured temperatures: north and south primary beams at midspan, and steel bolt temperatures at same height C1_4 & C4_4; (b). Thermocouple instrumentation locations at connection C1, tagged as C1_1 to C1_6; and (c). Thermocouple instrumentation locations at connection C4, tagged as C4_1 to C4_6.

The measured upper layer gas temperature, $\theta_{g,t}$, is used as the input variable of equation (3) with a 5 s time interval for calculation. Note that the measured gas temperature during the heating phase of the experiment, closely matching with the ISO-834 standard fire curve, provides a good benchmark to examine the applicability of Eurocode 3 since standard fire curves (e.g. the ISO-834 fire curve) are a common fire situation considered for design. In addition, this study utilized the steel member density ρ_a of 7850 kg/m^3, and the temperature-dependent specific heat of the steel c_a according to Eurocode 3. The section factor of the W16×31 shape, $A_p/V,$ was taken to be equal to 203 m^{-1} with a three-sided fire exposure. The SFRM thermal conductivity λ_p is 0.086 W/mK [20], and its specific heat $c_p$is assumed to be 1200 J/kgK [21]. There are also two SFRM-related parameters used for estimating uncertainties, including the unit mass, ρ_p, ranging from 240 kg/m^3 to 350 kg/m^3; and the applied thickness d_p. The measured value of d_p was 18 mm on average for the north primary beam, with 3 mm standard deviation and was 19 mm for the south primary beam with 2 mm standard deviation. The average temperatures of the north and south primary beams reached a maximum value of 780 °C and 670 °C, respectively, approximately 2 mins after the fire was extinguished (i.e., at 109 min). The Eurocode 3 prediction, incorporating the step-by-step method, on these primary beams suggests a maximum value ranging from[8] 580 °C to 690 °C at 115 min. The steel beam temperature predicted using the same method (equation (3)) appears to be lower than the corresponding values of measured temperatures, and yet the connection temperatures of the components C1_4 and C4_4 are conservative when calculated using equations (1) and (2). The predicted maximum temperature of those two connection components is approximately 540 °C, i.e. higher than their measured value of 400 °C on average.

The step-by-step section factor method using equation (3) is a preliminary step for calculating the steel connection component temperatures using equation (1) and (2). Therefore, it is worth investigating the accuracy of the section factor method in a more extensive manner, as presented in Figure 8 and further summarized in Table 2. For all steel beams, with an exception of the east primary beam, the predicted maximum temperatures of steel members are approximately 6% lower than the measured values on average.

[8] This range is due to the SFRM input property uncertainties: unit mass and thickness.

Furthermore, the average cooling rate[9] of all beams in the experiment is 110 ºC/hour, about 9% greater than the predicted rate. One possible reason for the discrepancy between the prediction and the measurement, is due to the fire conditions achieved during the experiment. In the experiment, natural gas was used as the fuel which seldom generated smoke, applying cumulative radiation to the surface of the passive fire protection; however, in a real building fire, this situation is highly unlikely. The sooty smoke within the compartment upper layer, generated from a real fire, would obscure some of the radiation from the flames to the fire protection.

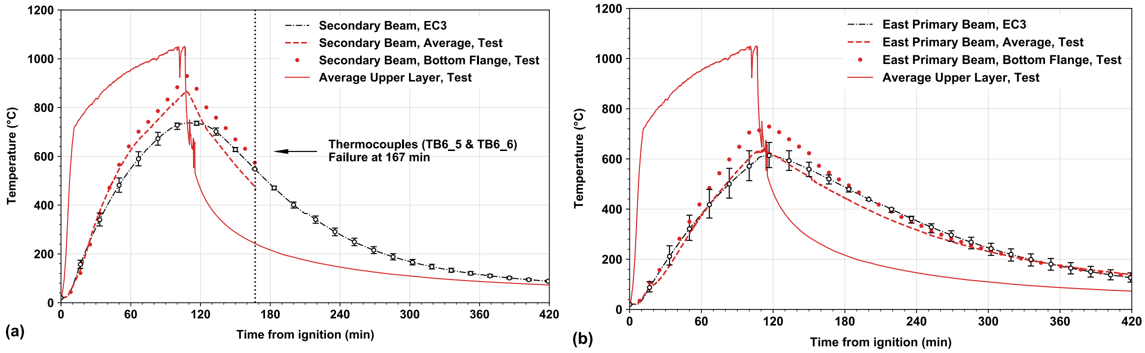


Figure 8. Comparison of the predicted with measured temperatures: (a). Secondary beam (thermocouples at the bottom flange of beam midspan, TB6_5 and TB6_6 failed at 167 min); and (b). East primary beam.

Table 2. Comparison between the test and the Eurocode 3 prediction on steel beam temperatures (for maximum temperature, Eurocode 3 method considers the upper bound prediction for comparison except for east primary beam; for cooling rate, Eurocode 3 method considers the mean value)

Beam at midspan	Maximum Temperature (ºC)		Cooling Rate (ºC/hour)	
	Test (Average)	Eurocode 3	Test (Average)	Eurocode 3
Secondary Beam	870	750	/	120
South Primary Beam	670	650	105	95
North Primary Beam	780	690	130	100
West Primary Beam	690	650	105	90
East Primary Beam	640	560 - 660	100	90
Average All Beams	730	680	110	100

Figure 9(a) and (b) present the predicted and measured values of maximum temperatures as well as cooling rates respectively, for all the measured connection components on the ten connections considered in this study. In Figure 9(a), the error bar of the test measured maximum temperature represents the standard deviation of two connection components on the same beam at two different ends; and the error bar of the Eurocode 3 predicted maximum temperature refers to the temperature variation due to the range of SFRM input variables, i.e. unit mass and thickness. Figure 9(a) suggests that Eurocode 3 tends to overpredict maximum temperatures of the connection components when actually heated to 400 ºC or lower. However, the predicted temperatures (using Eurocode 3) become comparable to the measured temperatures of the connection components when actually heated in excess of 400 ºC. It is important to repeat herein that this comparison is made under the situation where the SFRM thickness on the connection region was at least 43% thicker than the SFRM on the beams. To further examine the impact of SFRM thickness varying between the beam midspan and the connection region, Figure 9(a) also includes the comparison with

[9] The average cooling rate is calculated from the time of steel member at peak temperature, lasting five hours during the cooling phase.

another data from the long-span composite beam fire test carried out at NIST [22]. In this test, the SFRM thickness on the connection region was at least 68% greater than that on the steel beam. For this case, Eurocode 3 method overestimates the connection temperatures of which measured values were actually lower than 300 °C. Figure 9(b) explores the comparison between the measured and predicted cooling rates. It suggests that in most cases Eurocode 3 predicts much rapid cooling rates, from 70 °C/hour to 120 °C/hour for the connections used in this study, whereas the measured cooling rates significantly vary from 10 °C/hour to 110 °C/hour. This difference would be influenced by several factors, including the Eurocode 3 overestimation on maximum temperatures leading to a higher slope, the thicker SFRM applied on the connection region resulting in slow cool-down, or both.

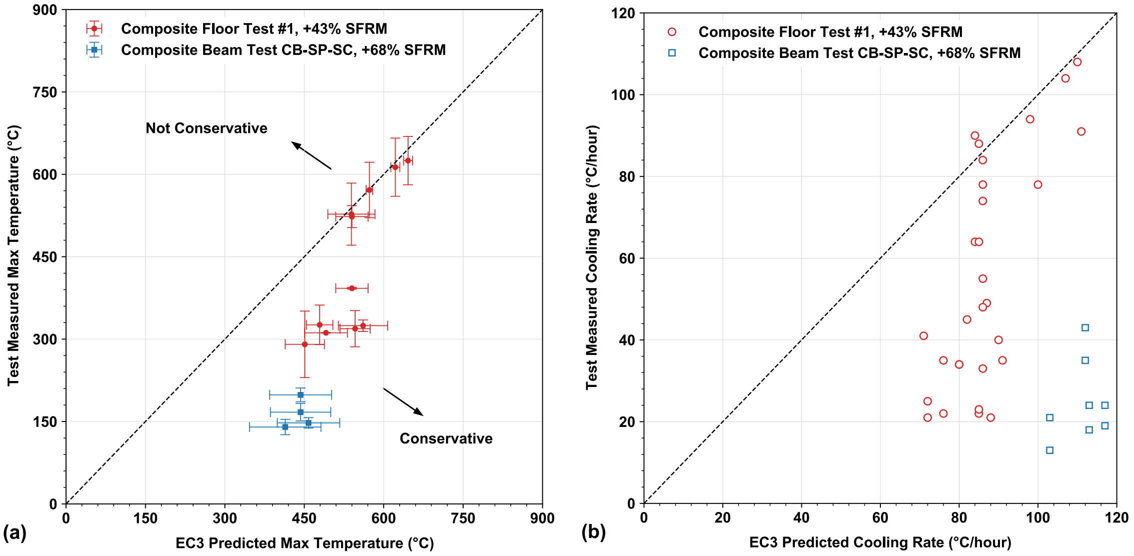


Figure 9. Comparison of the measured against Eurocode 3 predicted temperatures on all ten connections, i.e. 30 connection components in total plus another 8 connection components from composite beam test CB-SP-SC [22]: (a). Maximum temperature; and (b). Cooling rate within five hours.

4 CAPACITY OF THE CONNECTIONS

The shear capacity of welds and bolts used in the connection C2 was estimated using the Eurocode 3 reduction factors and experimentally measured temperatures. Figure 10(a) demonstrates that C2 would have failed at 575 °C (around 100 mins after the gas burner ignition). However, this behaviour was not witnessed during the experiment, see Figure 10(b) and (c).

Figure 11 illustrates the overestimation of connection temperature C2 predicted using the Eurocode 3 bolt strength reduction factor[10]. As shown, at 90 min, the bolt reduction factor decreases to as low as 0.11 when calculated using the measured bottom flange temperature of the steel beam at midspan or ranges from 0.20 to 0.31 when the Eurocode 3 step-by-step section factor method is used. Those two predictions are conservative, as compared to the actual 0.46 estimated via the temperature measurement at bolt C2_4.

However, in the case of C6 connection, the Eurocode 3 prediction of bolt strength reduction factor is less conservative, see Figure 12 at 90 min. This bolt reduction factor (based on the thermocouple measurement at this bolt) decreases to 0.53. This result is within the predictive range of 0.48 to 0.74 based upon the Eurocode 3 step-by-step section factor method but higher than the reduction factor 0.35 predicted using the measured beam bottom flange temperature at midspan. Although these predictions still imply no bolt failures, the post-fire inspection of the experiment discovered that three of five bolts from the connection C6 middle row to lower row experienced partial shear rupture failure, Figure 13. This was due to the

[10] Reduction factor: a ratio (≤1) between the steel bolt strength at high temperature and the strength at ambient temperature

combined large axial force and bending moment, induced by the thermal restraint and thermal bowing effects during the heating phase.

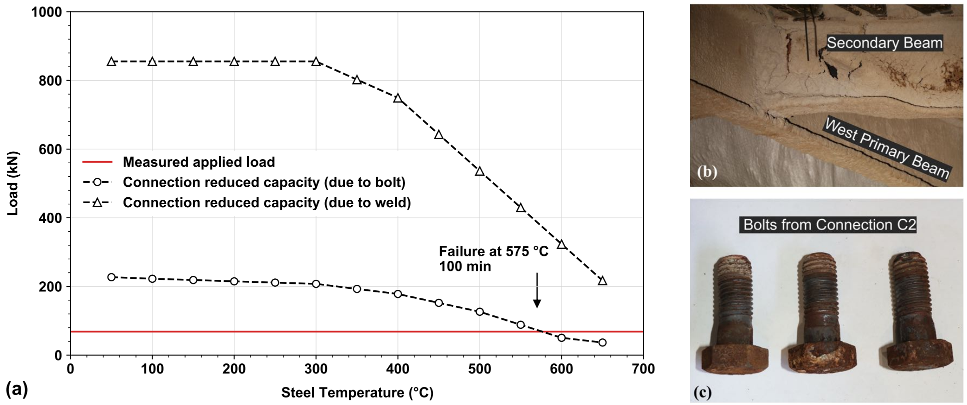


Figure 10. (a). Reduction of connection capacity with experimental increasing temperature at connection C2, against the corresponding measured applied load; (b). Post-fire inspection on connection C2 (before SFRM removed); and (c). Post-fire inspection on the bolts from connection C2.

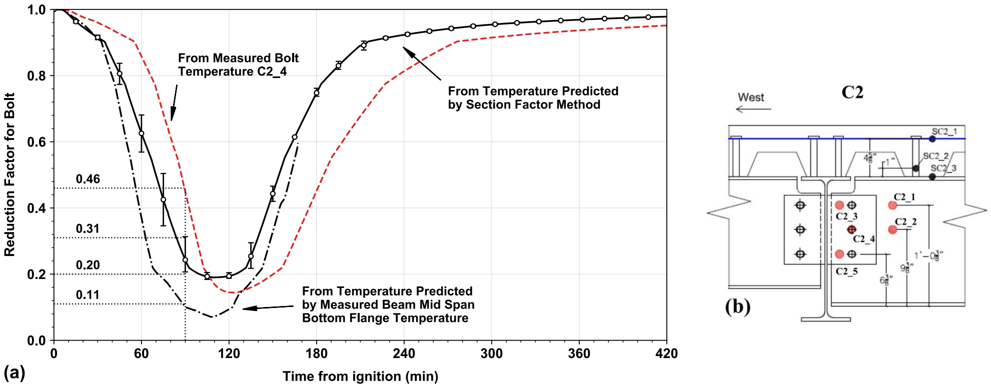


Figure 11. (a). Reduction factor of the bolt based on different methods and measured temperature at the C2 middle row, C2_4 (beam thermocouples at bottom flange failed at 167 min); and (b). Thermocouple instrumentation locations at connection C2, tagged as C2_1 to C2_5.

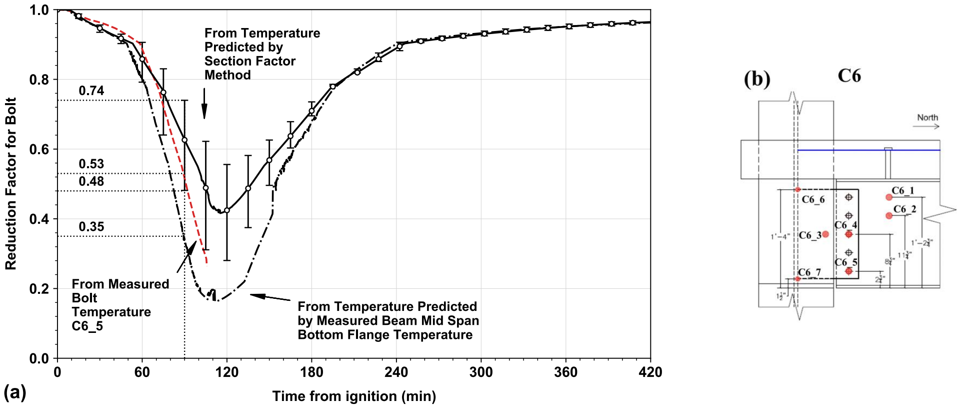


Figure 12. (a). Reduction factor of the bolt based on Eurocode 3 methods and measured temperature at the C6 lower row, C6_5 (bolt thermocouple failed at 106 min); and (b). Thermocouple instrumentation locations at connection C6, tagged as C6_1 to C6_7.

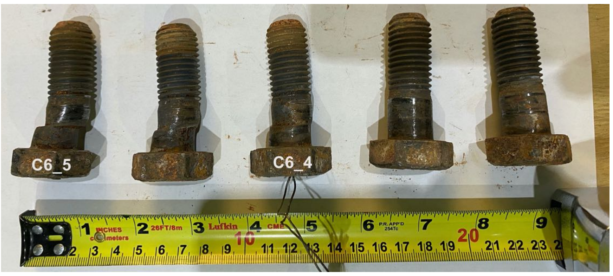


Figure 13. Post-fire inspection on the bolts from connection C6.

The findings in this section shows that Eurocode 3 method does not account for (1) additional sources constituting the capacity of a connection, such as the presence of slab and slab continuity to adjacent bays and (2) additional sources for the demand, such as thermally-induced axial forces due to the restraint to thermal expansion or contraction as well as catenary action. All these factors are needed to be incorporated for reliable estimation of the connection integrity.

5 CONCLUSIONS

The experimental results presented in this paper enrich the database that can be used for validation of computational models predicting beam end shear connections with fire protection under large compartment fires. The data acquisition of the measurements was successful, only 3 out of 45 thermocouple measurements on the total ten connection regions failed during the 7 hours (including cooling phase) of this structural fire test duration. This work demonstrates that the Eurocode 3 provision on the temperature prediction of connection components is conservative, provided that the fire protection at the connection region is at least 43% thicker than the protection on the beams. Finally, designing shear connections through temperature provisions may not guarantee a safe structural fire design. It is strongly recommended that the influence of the axial load demand (i.e. compressive/tensile load demands) of the connection *must also* be taken into account in future design guidance. This test indicated that the connection region can be subjected to varying axial forces during a fire event, such as a compressive force induced by the restraint of thermal expansion in the heating phase, followed by a tension force by catenary action and the contraction of beam members during cooling. The combined effects from high temperatures and fire-induced forces can lead to failure of connections designed only through Eurocode temperature provisions.

ACKNOWLEDGMENT

The support of numerous NIST colleagues on this project is acknowledged and greatly appreciated: Matthew Hoehler, Matthew Bundy, Selvarajah Ramesh, Brian Story, Tony Chakalis, and Philip Deardorff.

DISCLAIMER

Certain commercial entities, equipment, software, or materials may be identified in this paper 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. AISC, Specification for Structural Steel Buildings, ANSI / AISC 360-16. American Institute of Steel Construction (AISC), Chicago, Illinois, 2016.
2. Clifton, G. C., Meng, F., Mohammadjani, C. and Abu, A., Importance of Concrete Floor Slabs in Composite Beam to Column Connections during Severe Fires, ASFE, 2019, Singapore.

3. Chlouba, J., Wald, F. and Sokol, Z., Temperature of connections during fire on steel framed building, International Journal of Steel Structures, vol. 9, pp. 47–55, 2009. https://doi.org/10.1007/BF03249479
4. Fischer, C. E., Varma, H. A., Fire resilience of composite beams with simple connections: Parametric studies and design, Journal of Constructional Steel Research, vol. 128, pp. 119-135, 2017. https://doi.org/10.1016/j.jcsr.2016.08.004
5. Eurocode 3. Design of Steel Structures - Part 1-2: General rules — Structural fire design. European Standard EN 1993-1-2, CEN, Brussels, 2005.
6. Wang, Y., Burgess, I., Wald, F., Gillie, M., Performance-based Fire Engineering of Structures, CRC Press, Taylor & Francis Group, p363, 2013.
7. da Silva, L.S., Santiago, A., Real, P.V., Moore, D., Behaviour of steel connections under fire loading. Steel and Composite Structures. Vol. 5, No. 6, pp. 485-513, 2005. https://doi.org/10.12989/SCS.2005.5.6.485
8. Choe, L., Ramesh, S., Dai, X., Hoehler, M., Bundy, M., Experimental study on fire resistance of a full-scale composite floor assembly in a two-story steel framed building. Proceeding of the 11th International Conference on Structures in Fire (SiF' 20), Nov. 30 – Dec. 02, 2020, University of Queensland, Australia
9. ASCE, Minimum Design Loads and Associated Criteria for Buildings and Other Structures. ASCE/SEI 7-16, American Society of Civil Engineers, Reston, VA., 2016.
10. Southwest Type 5MD, Product Data Sheet, Carboline, 2019.
11. Underwriter Laboratory (UL), Fire Resistance Ratings – ANSI/UL 263. Design NO. N791, 2011
12. Underwriter Laboratory (UL), Fire Resistance Ratings – ANSI/UL 263. Design NO. D949, 2015
13. ASTM, Standard Specification for Carbon Structural Steel. ASTM A36/A36M - 19, ASTM International, West Conshohocken, PA., 2019.
14. ASTM, Standard Specification for High Strength Structural Bolts and Assemblies, Steel and Alloy Steel, Heat Treated, Inch Dimensions 120 ksi and 150 ksi Minimum Tensile Strength, and Metric Dimensions 830 MPa and 1040 MPa Minimum Tensile Strength1. ASTM F3125/F3125M - 19, ASTM International, West Conshohocken, PA., 2019.
15. Zhang, C., Grosshandler W., Sauca A., and Choe L., Design of an ASTM E119 fire environment in a large compartment, Fire Technology, pp. 1–23, 2019. https://doi.org/10.1007/s10694-019-00924-7
16. Sauca, A., Zhang, C., Grosshandler, W., Choe, L., and Bundy, M., Development of a Standard Fire Condition for a Large Compartment Floor Assembly, Technical Note (NIST TN) - 2070, pp 62, 2019. https://doi.org/10.6028/NIST.TN.2070
17. Bryant, R. and Bundy, M. The NIST 20 MW Calorimetry Measurement System for Large-Fire Research, Technical Note (NIST TN) - 2077, pp 68, 2019. https://doi.org/10.6028/NIST.TN.2077
18. ASTM, Standard Test Methods for Fire Tests of Building Construction and Materials. ASTM E119−19, ASTM International, West Conshohocken, PA., 2019.
19. Eurocode 1. Actions on Structures - Part 1-2: General Actions - Actions on Structures Exposed to Fire. European Standard EN 1991-1-2, CEN, Brussels, 2002.
20. Fire Protection Systems, Southwest Fireproofing Type 5 MD, 2008.
21. Franssen, J.M., Vila Real, P., Fire Design of Steel Structures, 2nd Edition, ECCS – European Convention for Constructional Steelwork, p450, 2016.
22. Choe L.Y., Ramesh S., Hoehler M.S., Seif M.S., Bundy M.F., Reilly J., Glisic B., Compartment Fire Experiments on Long-Span Composite-Beams with Simple Shear Connections Part 2: Test Results, Technical Note (NIST TN) - 2055, pp 144, 2019. https://doi.org/10.6028/NIST.TN.2055

Predicting Flashover Occurrence using Surrogate Temperature Data

Eugene Yujun Fu,[1,2,*] Wai Cheong Tam,[1,*,‡] Jun Wang,[1,2] Richard Peacock,[1]
Paul Reneke,[1] Grace Ngai,[2] Hong Va Leong,[2] Thomas Cleary[1]

[1]Fire Research Division, Engineering Laboratory, National Institute of Standards and Technology, Maryland, USA

[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
csyfu@comp.polyu.edu.hk, {waicheong.tam, jun.wang, richard.peacock, paul.reneke}@nist.gov,
{csgngai, cshleong}@comp.polyu.edu.hk, thomas.cleary@nist.gov

Abstract

Fire fighter fatalities and injuries in the U.S. remain too high and fire fighting too hazardous. Until now, fire fighters rely only on their experience to avoid life-threatening fire events, such as flashover. In this paper, we describe the development of a flashover prediction model which can be used to warn fire fighters before flashover occurs. Specifically, we consider the use of a fire simulation program to generate a set of synthetic data and an attention-based bidirectional long short-term memory to learn the complex relationships between temperature signals and flashover conditions. We first validate the fire simulation program with temperature measurements obtained from full-scale fire experiments. Then, we generate a set of synthetic temperature data which account for the realistic fire and vent opening conditions in a multi-compartment structure. Results show that our proposed method achieves promising performance for prediction of flashover even when temperature data is completely lost in the room of fire origin. It is believed that the flashover prediction model can facilitate the transformation of fire fighting tactics from traditional experience-based decision marking to data-driven decision marking and reduce fire fighter deaths and injuries.

Introduction

Fire fighters face tremendous dangers on the fire ground. Over the past ten years, nearly 750 fire fighters were killed and approximately 250,000 fire fighters were injured (Campbell et al. 2019; Fahy et al. 2020). Rapid fire progression, such as flashover, has been identified as one of the leading causes for both fire fighter fatalities and injuries. In a fire scenario, flashover is an extreme event. When it occurs, nearly all directly exposed combustible materials, such as a sofa, mattress, and carpeting, in a compartment, such as living room or bedroom, can be simultaneously ignited. Consequently, gas temperature within the compartment increases exponentially, and exceed more than 800 °C (Thomas et al. 1980). Figure 1a shows the compartment gas temperature profile from a flashover fire. In such a high temperature condition, survival for any fully-equipped fire fighter is rare (Dunn 2015).

Although typical indicators of the onset of flashover, such as hot layer gas temperature achieving approximately 550 °C to 600 °C (Peacock et al. 1999) and/or average heat flux at the floor level reaching 20 kW/m^2 to 25 kW/m^2 (Walton et al. 2016), are well known in the fire research community, this kind of detailed information about the interior thermal conditions is not available for nearly all fires. Thus, it is rather difficult for fire fighters to realize the potential fire hazards inside the fire room from outside.

In a structure fire, fire fighters rely on their experience in recognizing the potential occurrence of flashover. According to the most updated fire fighting training manual (Stowell and Murnane 2017), rollover is one possible flashover indicator. Visually, it can be seen as flames spreading across the ceiling outside of the fire room. When rollover phenomenon is observed, a flashover is likely to occur. However, this kind of experience-based indicator is not easy to recognize, and it could take many years of experience to build up the necessary proficiency. Therefore, if fire fighters do not have such a high level of situational awareness, the flashover threat presents itself as an unpredictable life-threatening hazard.

One can save a significant number of lives by developing a data-driven model based on temperature signals from heat sensors within the compartment to warn fire fighters before the flashover occurs. Yet, there are two primary challenges: (1) temperature data in multi-compartment structures are

[*] These two authors contributed equally.
‡ Corresponding author.

Tam, Andy; Fu, Eugene Yujun; Peacock, Richard D.; Reneke, Paul A.; Wang, Jun; Ngai, Grace; Leong, Hong Va; Cleary, Thomas. "Predicting Flashover Occurrence using Surrogate Temperature Data." Presented at 35th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI-21). February 02, 2021 - February 09, 2021.

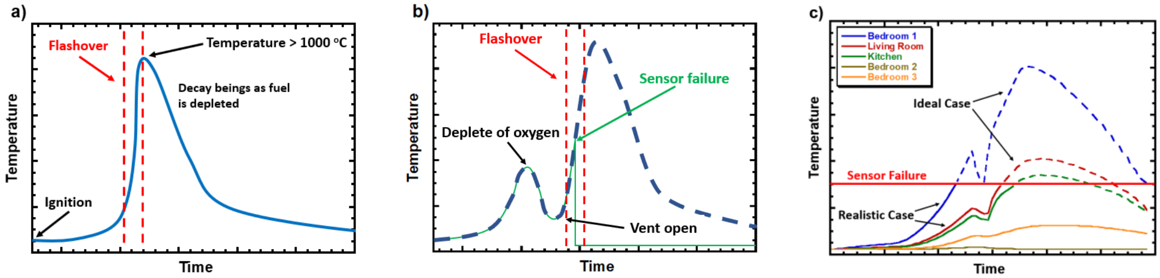


Figure 1: Temperature profile with flashover in a compartment from a) a fuel controlled fire and b) a ventilation controlled fire, and c) realistic and ideal temperature profiles for heat sensors at different compartments within a multi-compartment structure.

complex. In a typical fire scenario, no prior knowledge is given to the location of the fire, the item that is being ignited, and the opening conditions for interior and exterior vents (i.e., doors and windows). For example, a window breakage or a damaged door can create an opening from a closed room allowing oxygen-rich fresh air to enhance the fire. An example is given in Figure 1b. Given the right amount of fuel, oxygen, and heat (known as the fire triangle), flashover occurs.

What makes the problem more complicated is that existing fire protection devices, such as heat sensors being placed at various locations within the structure, are likely to be destroyed due to flame and/or elevated temperature (i.e., ~ 150 °C to 250 °C and see Figure 1b for sensor failure) (NFPA 2002). If there is no temperature signal in the room of fire origin (refer to the green solid line in Figure 1b), no direct prediction about the potential flashover occurrence can be made. Although the remaining temperature signals from other compartments can be used as surrogates, since the interior opening conditions are unknown and it is not clear which temperature signals are useful (see the solid lines in Figure 1c in which the temperature variation across different compartments is substantial), the prediction will become highly uncertain. To the best of the authors' knowledge, no empirical expressions nor models exist that can efficiently correlate relationships between temperature from non-fire rooms and flashover in multi-compartment structures.

The second challenge is that temperature data in real fires for full-scale multi-compartment structures is limited. Firstly, it can be easily understood that flashover does not frequently happen (Ahren 2019). Even if it happens, data is difficult to collect, and the data quality is questionable because important information such as exact fire location, ignited items, and/or vent opening conditions might not be well documented in case of a fire accident. Secondly, temperature data associated with flashover in building structures are not available from any public data repository (i.e., Dua and Graff 2019). Lastly, physically conducting full-scale fire experiments involving flashover in a multi-compartment structure with living room, dining room, kitchen, bedrooms, doors, and windows is extremely costly and time-consuming. Given the data limitation and the numerical bottleneck, the development of a data-driven model for the prediction of potential flashover occurrence accounting for the realistic effect of fires and vent opening conditions in multi-compartment structures requires innovative approaches.

In this paper, we propose the utilization of the machine learning paradigm with the learning by synthesis approach to overcome these challenges. The main contributions of this work are summarized as follow:

- **Problem**: We engage a novel problem about fire fighting accounting for realistic fire and vent opening conditions in a multi-compartment structure. We propose to develop a flashover prediction model (P-Flash) which can be used as a potential solution to transform traditional fire fighting tactics from experience-based decision making to data-driven decision making to enhance situational awareness, operational effectiveness, and safety for fire fighting and to enable smart fire fighting (Hamins et al. 2015).
- **Algorithm**: We propose to use an attention-based bidirectional long short term memory to capture crucial relationships between temperature data and flashover conditions. The model is able to differentiate temperature information with higher significance and provide flashover prediction even when the temperature signal from the room of fire origin is completely lost.
- **Data**: We provide 5041 sets of synthetic temperature data accounting for fire scenarios with a wide range of fire and vent opening conditions within a single story residential building. The core difference between our data generation process and that found in other literature is that our fire simulation program is validated against real-life experimental data with identical settings. This validation process helps to ensure the reliability of our synthetic data. Data and code are available upon request.
- **Evaluation**: We evaluate P-Flash against real data obtained from 13 different full-scale fire experiments with the occurrence of flashover (Madrzykowski and Weinschenk 2019). Experimental results reveal that our

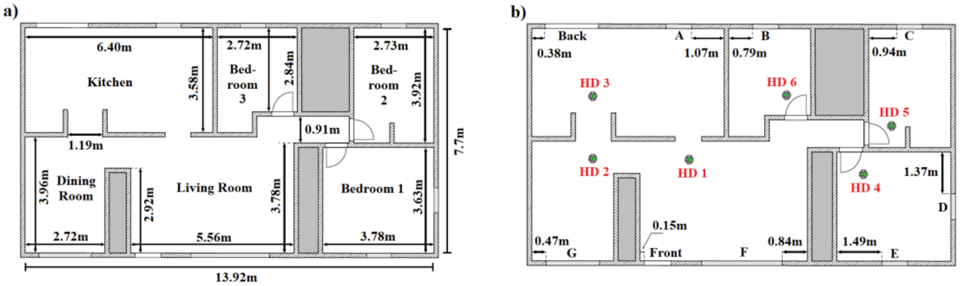


Figure 2: Plan view dimensioned drawing of a) the single story structure and b) vent openings with heat sensors (HD).

proposed method is feasible and hence has potential impact to real-world fire fighting.

Related Work

Flashover Prediction Models: Due to the technical complexity associated with the collection of temperature measurement for flashover conditions in full-scale experiments, research efforts primarily focus on single compartment structures. In the fire research community, correlation techniques relating air temperature and heat release rate (HRR) are typically being used for the estimation of flashover (Babrauskas 1980; McCaffrey et al. 1981; Deal and Beyler 1990; Richards et al. 1997; Overholt and Ezekoye 2012). The HRR can be understood as the rate of heat generation by a fire. Given an estimated HRR, the occurrence of flashover can be approximated. However, since these models are developed based on data obtained from small single compartments with approximately 16 m^2 in floor area and a single door-like vent, these models have limited applicability to multi-compartment structures.

Additional efforts are made to account for the geometric effect of flashover conditions (Yu et al. 2012; Zhang et al. 2014; Li et al. 2019; Kurzawski and Ezekoye 2020). Although their research outcomes provide substantial improvement for the development of flashover prediction models in multi-compartment structures, their models rely on assumptions that over-simplify the fire scenarios. Specifically, (i) all interior and exterior openings, such as doors and windows, are always assumed to be fully opened; (ii) fire locations are assumed to be at only one location; (iii) fire growth of burning items (i.e., how fast and how intensive the item is being combusted) is prescribed based on arbitrary functions without experimental validations; and (iv) most importantly, sensors being used to obtain the temperature signals are assumed to be ideal, meaning that the sensors will never fail. In contrast to the previous works, the realistic conditions involving items (i) – (iv) are considered in this present study. By doing so, our flashover prediction model is more suitable to provide flashover warnings to fire fighters for fire fighting in multi-compartments structures.

Generation of Synthetic Dataset using Simulation Models: The idea of using fire simulation models to generate synthetic data have been shown useful in recent studies such as fire detection in tunnel (Wu et al. 2020), structural fire protection design (Zhang et al. 2020), and hazard assessment (Lattimer et al. 2020) as it avoids the need of conducting costly experiments and facilitate parametric studies of a problem. For example, Wu and his coworkers (Wu et al. 2020) used a CFD model to generate detailed smoke and temperature data for different heat sensors at various locations with a wide range of fire and wind conditions. The advantage is clear. However, one potential concern is that they had never benchmarked and/or validated the model against full-scale experiments with similar fire and wind conditions. For that, it is uncertain if the synthetic data being generated from the model could capture the fire behavior correctly. In contrast, the reliability of our synthetic data is assured. Specifically, the synthetic data generated by our fire simulation model are benchmarked against experimental data (McKinnon et al. 2020) with identical settings. Therefore, we can quantify the accuracy of the fire simulation model in simulating the corresponding behaviors of fire and vent openings in a multi-compartment structure.

Multivariate Time Series Classification: Recurrent neural network (RNN) approaches, such as long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997), have achieved much success for various tasks in different scientific communities such as detection of mechanical failure (Guo et al. 2017), hurricane trajectory prediction (Alemany et al. 2019), understanding human communications (Zedeh et al. 2019), and early fake news detection (Liu and Wu 2019). Although LSTM is an efficient way to encode multivariate time series data, it processes inputs in temporal order in which its outputs tend to be mostly based on previous information without making full use of available information (Graves and Schmidhuber 2005). For the development of a flashover prediction model for situations similar to that of shown in Figure 1b, making use of all available information is crucial. Moreover, the standard LSTM may not have capabilities to discriminate data with higher significance, such as those temperature profiles (i.e., solid lines in red and

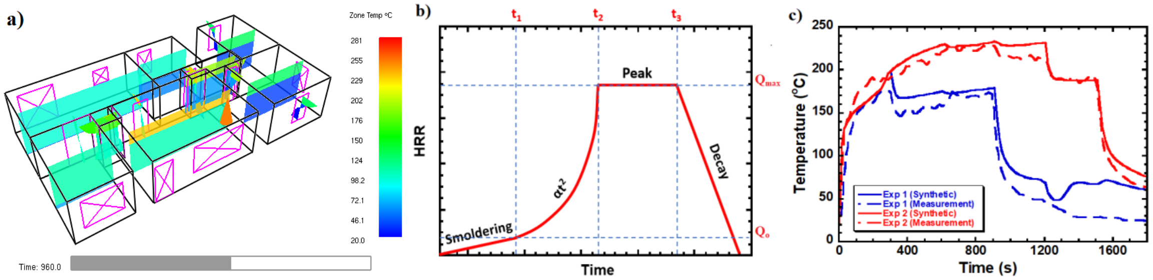


Figure 3: a) overview of a CFAST simulation run with a fire in the living room, b) standard t-squared fire HRR curve, and c) validation for Experiment 1 and 2 between CFAST results and measurements.

green) in Figure 1c. In this study, we will use the state of the art RNN architecture, namely bidirectional long short-term memory (BiLSTM) (Graves and Schmidhuber 2005) together with attention mechanism (Vaswani et al. 2017) to facilitate the learning of inherent patterns and complex relationships between temperature signals from non-fire compartments and flashover from the fire origin with realistic fire scenarios and arbitrary vent opening conditions.

Flashover in a Multi-Compartment Structure

Consider a single-story ranch structure as shown in Figure 2a. There are six different compartments: a living room, a dining room, a kitchen, and three bedrooms. The overall interior dimensions of the structure are roughly 13.92 m x 7.7 m with a ceiling height of 2.44 m. The detailed dimensions associated with each of the compartments are illustrated in Figure 2a. Since fire rarely occurs in bathrooms, bathrooms are not considered in current layout. For interior finish, the walls and ceiling are covered with gypsum wallboards and the floor is covered by cement board.

Figure 2b shows the relative position of vents and heat sensors in different compartments. For vents, there are two exterior doors (front and back), three bedrooms doors, a doorway that leads to the kitchen, and seven windows (A - G). For heat sensors, one heat sensor is located at each compartment, and they are about 0.02 m away from the ceiling. It is worth noting that this single story, traditional ranch style structure is selected because 90 % of residential buildings were built using this layout in the mid of 1950s (Madrzykowski and Weinschenk 2019). Currently, this structure remains the most popular style of home in 34 states across the United States (Mattern 2017). For that, the flashover prediction model developed based on this structure is expected to have substantial benefits for fire fighting across the U.S. Additional efforts accounting for the effect associated with different structure layouts is underway. Findings will be reported in future studies.

Synthetic Data Collection: CData (Tam et al. 2020) is utilized to execute simulation runs to generate the synthetic temperature data for a single item ignition fire with a wide range of fire and vent conditions. In general, CData is a Monte Carlo based sampler that uses CFAST[2] as the simulation engine. In this study, 5041 set of cases are considered, and it is consisted of about 1 million data points.

Three realistic conditions are taken into account in the data generation process, and they are 1) experimental validated fire growth of single burning items, 2) various fire locations, and 3) arbitrary opening conditions of vents.

Realistic Fire Growth of a Burning Item: Heat release rate (HRR) is the single most important variable in characterizing the fire growth of an item (Babrauskas and Peacock 1991). In order to obtain the HRR, experiments are typically performed, and Figure 3b shows the standard HRR curve. Specifically, this is the t-squared HRR curve that is used to describe the overall burning behavior of a single item in the fire research community. As shown in the figure, a burning item might experience four different fire growth stages: smoldering, t-squared growth, peak, and decay. Flashover usually happens in approximately between the t-squared growing stage and the peak stage. For that, experimentally validated HRR curves are crucial in capturing the precise burning behavior of an item such that the corresponding fire growth can be matched closely to actual fire scenarios. Four items, including a flaming chair, smoldering chair, polyurethane foam mattress, and cotton based mattress, are considered in this study (Reneke et al. 2019). The selection of these items is due to the fact that these items represent the largest portion of first ignited item in home fires (Ahrens 2017).

[2] CFAST (Peacock et al. 2015) is a fire simulation program that divides compartments into two zones. Each zone includes a gas mixture/soot medium bounded by a ceiling or a floor, and four surfaces. Conditions of each zone are assumed to be uniform. When there is a fire, a hot layer will form and the medium can be divided into an upper layer and a lower layer. If the fire persists, the upper layer increases in depth and the temperature will rise. When openings exist, there will be natural flow through the openings allowing air exchange between different compartments and zones. Figure 3a shows a simulation case for the single-story ranch structure with a fire in the living room.

Table 1 provides the table for the summary of HRR parameters associated with the four different items.

Fire location and Vent Openings: A fire can be initiated at the center of either one of the six different compartments in each simulation. Since the fire simulation model being used is a zone model, the exact location of the fire does not have any significant impacts to the resulting temperature (if the fire is not attached to any walls or corners). In the current dataset, the number of fire cases is distributed evenly for the six different compartments (i.e., about 840 cases for each compartment). For vent openings, all doors and windows within the structure, except the front door, are randomly selected to be either opened or closed at the beginning of a simulation run. For the front door, it can be opened at any time during a run. This arrangement accounts for the effect of different opening vents. In this current study, each of the vents is assigned to be opened for 60 % of the total cases. The value is chosen because we want to facilitate flashover conditions.

Validation with Experimental Data: In order to make sure that CData can be used to generate realistic temperature data for different fire scenarios, validation is carried out. Specifically, temperature measurements obtained from two full-scale experiments reported in (McKinnon et al. 2020) with a fire initiated in the living room within the single story residential structure are used to benchmark the synthetic data. The fire location and the HRR of the burning item for the two tests are the same. Yet, opening conditions of each of the vents are different. The details of the opening time for each vent is provided in table attached to Table 2. It is worth noting that natural gas burners are used in these experiments. The reason is that the HRR of the fire can be fully controlled by regulating how much natural gas is being burned. By doing so, we can be assured that the simulation conditions and the experimental conditions are identical.

Figure 3c shows the temperature measurements (dash lines) and the synthetic temperature data (solid lines) from the living room sensor for the two experiments. The overall agreement is great. It can be seen that the magnitude and trend of the temperature profiles matches the experimental data for different vent opening events. This observation indicates that CFAST, the simulation engine of CData, is capable of capturing both the corresponding effect of fire and vent openings in the single story multi-compartment structure. In terms of uncertainty, the absolute root mean squared error is about 30 °C and 10 °C for Exp 1 and Exp 2, respectively. Therefore, it can be said that the generated data is reliable.

Algorithm

Given the synthetic set of temperature data, our model will have to be able to carry out the following two tasks: 1) to relate complex data behavior to flashover conditions accounting for the effect of different fire and vent opening conditions and 2) to discriminate data with higher significance (see Figure 1c) and encode contextual information.

Table 1: HRR parameters (Reneke et al. 2019; Kim and Lilley 2002).

Items	Q_o (kW)	Q_{max} (kW)	t_1 (s)	t_2-t_1 (s)	t_3-t_2 (s)
Flaming Chair	10 - 30	270 - 3500	150 – 1250	90 - 600	200 – 400
Smoldering Chair	10 - 20	250 – 2500	5000 – 11000	70 - 500	200 – 400
Mattress (foam)	20 – 55	2200 – 4700	150 – 1250	200 -600	150 – 300
Mattress (cotton)	15 - 40	150 – 820	150 – 1250	30 – 1400	250 – 550

Temperature Signal Learning: In order to overcome the challenge associated with the 1st task, we propose the use of bidirectional long short-term memory (BiLSTM) (Graves and Schmidhuber 2005). Figure 4a shows the overall model architecture. It can be seen that for a temperature signal: $S = (s_1, s_2, \dots, s_\tau)$ and a time step i, BiLSTM includes a forward hidden state $\overrightarrow{h_i}$ and a backward hidden state $\overleftarrow{h_i}$. In this study, since we are interested to capture the complete behavior for temperature signals, we only make use of the last hidden state of $\overrightarrow{h_\tau}$ and $\overleftarrow{h_\tau}$. As shown in the figure, concatenation is applied to yield $h_\tau = [\overrightarrow{h_\tau}, \overleftarrow{h_\tau}]$ to encode temperature behavior with flashover conditions.

Sensor-Wise Self-Attention: In order to enhance the learning capability of the model in discriminating temperature signals with higher significance for more reliable prediction (i.e., neglecting bedroom 2 temperature signal in Figure 1c), we utilize a self-attention mechanism to model sensor-wise relation. For that, we will be able to extract the contextual temperature information of all compartments within the structure. It is believed that the contextual information can contribute to provide more accurate flashover predictions.

As shown in Figure 4b, our model takes the temperature signals from all compartments ($\{S_1, \dots, S_N\}$, $N = 6$) as inputs. Using the BiLSTM, we obtain the hidden state of temperature behavior (h_τ) for each signal. We then feed them into a sensor-wise self-attention module for sensor relation

Table 2: Event sequence.

Event	Exp 1	Exp 2
Front Door Open	300 s	1200 s
Back Door Open	1275 s	1860 s
Window A Open	1260 s	1845 s
Window B Open	1245 s	1830 s
Window C Open	1230 s	900 s
Window D Open	1215 s	Closed
Window E Open	600 s	600 s
Window F Open	1200 s	1815 s

modeling. Specifically, the attention weight (α_{ij}) of each pair of sensor signals (S_i, S_j) is determined based on the interaction of their modeled temperature behavior (h_τ^i, h_τ^j):

$$a_{ij} = {h_\tau^i}^T h_\tau^j \tag{1}$$

$$\alpha_{ij} = \frac{\exp\,(a_{ij})}{\sum_{k=1}^{N} \exp\,(a_{ik})} \tag{2}$$

To obtain the contextual temperature information that captures the temperature behaviors of all compartments. We extract context features based on the learned attention weights. For a signal S_i, we compute its context feature as:

$$e_i = \sigma(W_e h_\tau^i + b_e) \tag{3}$$

$$c_i = \sum_{j=1}^{N} \alpha_{ij} e_j \tag{4}$$

where σ is the activation function, and W_e and b_e are the parameters of a dense layer for further encoding temperature behavior. We attain the final feature representation u_i' for S_i after applying one dense layer on the concatenation of its context feature and encoded feature: $u_i = [c_i, e_i]$ such that:

$$u_i' = \sigma(W_c u_i + b_c) \tag{5}$$

We compute the representation for all temperature signals by the same manner to acquire the overall representation of the whole structure: $\mu' = [u_1', \dots, u_N']$. This feature representation is used to predict whether there is a flashover occurrence within the coming x seconds based on the available temperature signals.

Evaluation

Experimental Settings: Each synthetic fire experiment has six temperature signals (S) and each signal is corresponding to a compartment (i). The temperature signals from the compartment are denoted as $S^i = (s_0^i, s_{15}^i, \dots, s_T^i)$ where s_0^i and s_T^i are the first and the last temperature for an experiment, respectively, and T is the total duration. The sampling interval for all temperature signals is 15 s and this is selected to facilitate the data generation process. In total, there are 5041 synthetic fire experiments/events.

Sliding window is applied and instances are constructed. An instance from a fire event is formulated as $I_k = \{S_k^1, \dots, S_k^6\}$ where $S_k^i = (s_k^i, \dots, s_{k+w}^i)$ with k to be the first time step of the sliding window and w to be the window size. Accounting for the sensor limit, we adopt the sensor failure threshold of 250 °C from (NFPA 2002). Given the threshold, the sensor failure moment (T_b^i) for signal S^i can then be determined. If time $t \geq T_b^i$, s_t^i is replaced by a value of 0 °C, representing a loss of sensor signal. A masking layer is applied to neglect the zero values. Extracting all the I_k from all fire events, the instance set $\{I_0^1, \dots, I_k^{5041}, \dots\}$ is obtained.

Our task is to predict whether flashover will occur within the next x seconds based on the temperature data in I_k^e. In our experiment, we evaluate the models when $x = 30$ s and $x = 60$ s. These values are chosen with careful consideration about the response time in actual fire fighting (Dunn 2015). Due to movement limit (i.e., crawling to avoid excessive heat from ceiling), it will take 10 s for fire fighters to travel for approximately 3 m in a fire scene. For that, predictions ahead of flashover occurrence is crucial in order to allow the fire fighters to get away from the dangerous compartments or find shelters. Therefore, we will examine our model performance for $x = 30$ s and $x = 60$ s.

Each instance is labeled to form our data samples, and the instance is either labeled as Flashover or Non-Flashover based on its future temperature value. In our study, we take 550℃ to be the threshold of the onset of flashover conditions. In the current dataset, we have data imbalance for samples associated with Flashover and Non-Flashover. It can be understood that when $x = 30$ s, we only have two Flashover samples in one fire event (i.e., I_{f-15-w}^e and I_{f-30-w}^e where f is the flashover moment for event e). And when $x = 60$ s, the number of Flashover samples is four for one fire event (i.e., I_{f-15-w}^e, …, I_{f-60-w}^e). However, there are many Non-Flashover samples. In order to overcome the data problem, we first take all the Flashover samples and randomly select two (four) Non-Flashover samples for $x =$

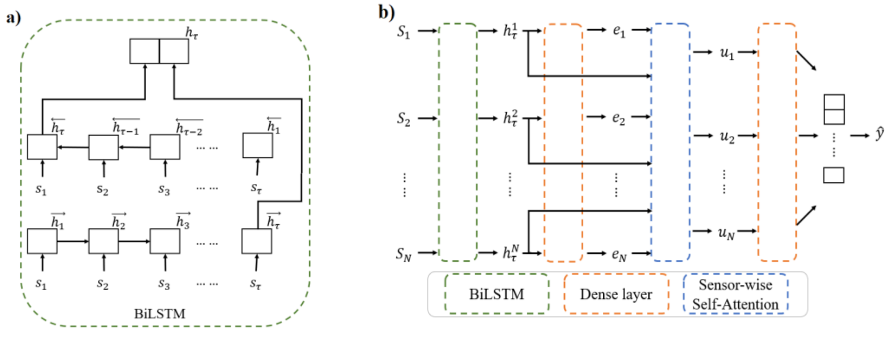


Figure 4: a) Model architecture of BiLSTM and b) BiLSTM with sensor-wise self-attention.

Table 3: Performance of flashover prediction.

x	Model	Acc.	Prec.	Rec.	F1
60s	BiLSTM	81.80%	86.88%	74.90%	80.45%
	BiLSTM-Attention	86.46%	84.54%	89.24%	86.82%
30s	BiLSTM	78.17%	76.94%	80.46%	78.66%
	BiLSTM-Attention	81.75%	79.47%	85.62%	82.43%

30s (60s) from each fire event. These samples are used to form our final dataset. In this experiment, we have 20164 and 40328 data samples (from 5041 fire events) for the experiment of adopting x as 30s and 60s, respectively.

For training and testing, we randomly split the data samples to form subsets for training, validation, and testing based on the fire events. Specifically, a set of 504 fire events worth of data samples are assigned to both validation set and testing set, respectively. The data samples from the rest of the 4033 (5041 – 2*504) fire events are given to the training set. The data proportion for training and testing process are identical to both $x = 30$ s and x = 60 s.

Model Configurations: We set the dimension of both the forward and backward LSTM as 28. The output dimension of our BiLSTM module is 56. We adopt 28 and 16 as the output dimension of the first and second dense layer in our model. Dropout is applied in the network with a dropout rate of 0.2. In order to evaluate the efficiency of the proposed sensor-wise self-attention module, we compare the performance of the BiLSTM model with attention (refer to it as BiLSTM-Attention) and the BiLSTM model without attention (refer to it as BiLSTM) in our experiments.

Experimental Results: Table 3 shows the model performance for flashover prediction made in 30 s and 60 s. Based on the accuracy and F1 scores, it can be shown that the proposed attention-based BiLSTM outperforms the original BiLSTM. It is worth noting that the attention-based model also yields a significantly better recall score, indicating the benefits of including the sensor-wise self-attention mechanism. This is extremely important for life saving purpose in fire events.

As shown in Table 3, the overall performance for prediction of flashover occurrence with x = 60 s is generally better. One possible reason is due to the fact that it has a larger set of training samples. For future work, we will keep collecting data to train a better prediction model for both scenarios.

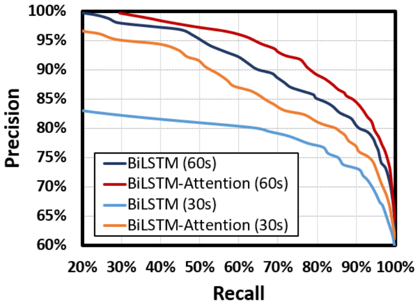


Figure 5: Precision and recall curves.

Precision-recall curves are illustrated in Figure 5. In real-life application, we would like to maximize the value of true positive for prediction of flashover occurrence with minimal or even zero false positive to avoid disturbance to fire fighters. As shown in Figure 5, it is observed that the attention-based model can obtain a precision value of approximately 85% with the recall value being above 90% (for 60 s). Generally, our proposed attention-based models (both 60 s and 30 s) can achieve higher recall with higher precision. The overall performance of the attention-based model is therefore more robust than those without the attention mechanism. The success of the attention-based model is built upon its ability of determining the relation of different sensor signals for different vent opening conditions.

Figure 6 illustrates the learned attention weights between sensor signals in fire origin room and other compartments: kitchen (K), dining room (D), living room (L), and bedroom 1 to 3 (B1, B2, B3), for two door opening conditions: all opened denoted as Open and all closed denoted as Close. Our attention-based model can discover the spatial relation between sensor signals from different compartments. For instance, when fire occurs in kitchen, the signal of dining room and living room are determined as the most discriminating surrogate signals by the model (Figure 6a). And those from dining room and kitchen are taken as the most useful surrogate signals, when it comes to predicting flashover in living room (Figure 6b). These consist of the spatial relations of dining room, kitchen, and living room. The sensors placed in these three rooms are very close to each other. Hence, the model can predict flashover in one of them via modeling signals from the other twos, even when the doors are closed. On the other hand, the signals from dining room and kitchen are barely important for our model when fire occurs in bedroom 1, regardless of the door opening conditions (Figure 6c). This also agrees with their spatial relations

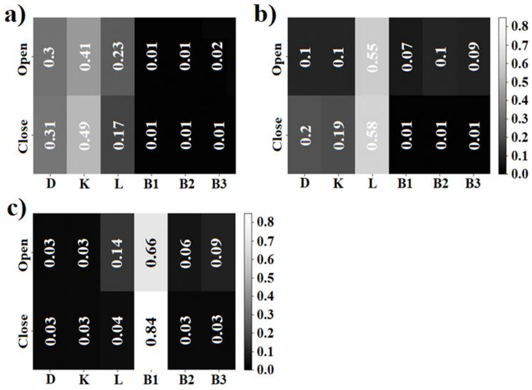


Figure 4: Learned attention for (a) Kitchen, (b) Living room, and (c) Bedroom 1, under different door opening conditions.

Tam, Andy; Fu, Eugene Yujun; Peacock, Richard D.; Reneke, Paul A.; Wang, Jun; Ngai, Grace; Leong, Hong Va; Cleary, Thomas. "Predicting Flashover Occurrence using Surrogate Temperature Data." Presented at 35th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI-21). February 02, 2021 - February 09, 2021.

Table 4: Key information for each of the experiments.

Exp #	Fire Location	Ignited item	Ventilation
1	Living Room	Sofa	All Vent Closed
2	Living Room	Sofa	All Vent Closed
3	Living Room	Sofa	Front Door Open
4	Living Room	Sofa	Front Door Open
5	Living Room	Sofa	Front Door and Bedroom 3 Window Open
6	Kitchen	Cabinet	All Vent Closed
8	Kitchen	Cabinet	All Vent Closed
10	Kitchen	Cabinet	Front Door Open
11	Kitchen	Cabinet	Front Door Open
7	Bedroom 1	Mattress	All Vent Closed
9	Bedroom 1	Mattress	All Vent Closed
12	Bedroom 1	Mattress	Front Door and Bedroom 1 Window Open
13	Bedroom 1	Mattress	Front Door and Bedroom 1 Window Open

that the sensors in dining room and kitchen are farer away from that in bedroom 1.

Moreover, different door opening conditions may also influence the relations of the sensor signals. Our attention-based model can also capture that. For example, our model determines that signal in living room has stronger relationships with that of bedroom 1, 2 and 3 when all the doors are opened, compared to the situation of closing all the doors (Figure 6b). Also, our model can find that for predicting flashover in bedroom 1, signals of living room, bedroom 2 and 3 are more useful when all the doors are opened than closed (Figure 6c).

The learned attention weights indicate that our attention-based model can successfully determine the relationships between different sensor signals under different door opening conditions. This results in extracting the most discriminating signals and contextual information. Based on that, a more reliable flashover prediction can be provided.

Towards Prediction in Real Fire Event: Given a flashover prediction model trained based on synthetic data (BiLSTM-Attention with 30 s), it is necessary to examine its performance against real-life fire scenarios. In this evaluation process, 13 sets of full-scale experiments reported in (Madrzykowski and Weinschenk 2019) are utilized. The building structure is identical to that of shown in Figure 2a. In these experiments, a single item is first ignited in either living room, kitchen, or bedroom 1. Temperature measurements are obtained and videos from thermal image cameras are recorded. Important information for each of the experiments is summarized in Table 4. As shown in the table, although a number of tests are repeated, the fire growth is rather different. It should be noted that our model never sees any of the experimental data.

Table 5: Model performance against real data.

	Instance I	Instance II	Non-Flashover
All	64%	54%	89%
Living room	80%	60%	80%
Kitchen	100%	100%	100%
Bedroom 1	25%	0%	100%

Due to the nature of the experiments, data associated with flashover and non-flashover conditions are imbalanced. In order to provide a fair comparison, the model performance is assessed based on 4 selected instances. With that, the model will make predictions about 15 seconds (Instance I) and 30 seconds (Instance II) prior to the flashover occurrence. These two instances are labeled as true for flashover occurrence. The exact time of flashover is obtained from the experimental data. The Instance III and IV are the non-flashover instances.

Table 5 shows the prediction accuracy for the instances associated with all 13 experiments and the 3 individual test series. It can be seen that the model performance for kitchen fire and bedroom 1 fire associated with either Instance I or II is substantially different. Fundamentally, this deficiency is primarily due to the fact that the HRR curves obtained based on standard experiments do not account for the enclosure effect (Ramesh and Venkateshan 1999). In principle, the high temperature environment surrounding the ignited item will enhance its pyrolysis process (Garrido and Font 2015), increasing the release rate of combustible gases, and accelerate the burning of an item. This combustion process is not being captured in current experiments[3] for HRR determination. For that, although experimentally validated HRR curves for foam mattresses is used during the data generation process in this study, the corresponding temperature behavior is substantially different. The rate of increase of temperature observed from the full-scale experiments is nearly double as compared to that of seen in our training data. For the kitchen fire, since wooden materials are the primary burning items, they have less influence with the enclosure effect. For that, the experimental temperature data are within the range of our synthetic data. Therefore, the flashover prediction is excellent. This observation is encouraging, and this is because a reliable flashover prediction model can be built even with synthetic temperature data if correct HRR curves accounting for the enclosure effect are applied. Also, it is believed that when the new data is available, the model performance for the flashover prediction model can be greatly enhanced.

Conclusion

In this paper, we present the development of a flashover prediction model for a multi-compartment structure using an attention-based BiLSTM with validated synthetic temperature data. This is the first work in which the realistic effects of fire locations, burning behavior of ignited items, vent

[3] For standard experiments, an item is ignited in a room temperature environment. Therefore, the pyrolysis process is sustained due to its own combustion and the enclosure effect due to room temperature is relatively small.

opening conditions, and data limitation due to sensor failure are being accounted for at the same time. Our model achieves promising performance. For synthetic datasets, it has the accuracy of ~ 86 % and ~ 82 % with the F1 score of ~ 87 % and ~ 82 % for prediction of flashover occurrence within the next 60 s and 30 s, respectively. The model performance is also tested against real data with flashover conditions in full-scale fire experiments. The overall accuracy for prediction of flashover occurrence is ~ 75 %. In the future, we will carry out physical experiments to account for the enclosure effect in HRR determination. Also, we are interested in developing a more generic flashover prediction model that can be used in any single story structure layout. It is believed that the flashover prediction model can help to save lives by enhancing situational awareness, operational effectiveness, and safety for fire fighting and enable smart fire fighting.

Acknowledgments

The authors would like to acknowledge Dr. Daniel Madrzykowski and Dr. Craig Weinschenk from the Underwriters Laboratories Firefighter Safety Research Institute (FSRI) for providing the valuable full-scale experimental data. Many thanks are due to Dr. Michael Xuelin Huang for helpful discussions.

References

Alemany, S., Beltran, J., Perez, A., and Ganzfried, S. 2019. Predicting Hurricane Trajectories using a Recurrent Neural Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ahrens, M., 2017. *Home fires that began with upholstered furniture*. National Fire Protection Association. Quincy, MA.

Ahrens, M. 2019. Home structure fires, Report, National Fire Protection Association, Quincy, MA.

Babrauskas, V. 1980. Estimating Room Flashover Potential. *Fire Technology*, 16(2): 94-103.

Babrauskas, V. and Peacock, R.D. 1992. Heat release rate: the single most important variable in fire hazard. *Fire safety Journal* 18(3): 255-272.

Campbell, R., Evarts, B., and Molis, J.L. 2019. Firefighter Injuries in the United States – 2018, Report, National Fire Protection Association, Quincy, MA.

Chen, J., Zhang, X., Zhao, Y., Bi, Y., Li, C., and Lu, S., 2019. Oxygen Concentration Effects on the Burning Behavior of Small Scale Pool Fires. *Fuel* 247: 378-385.

Deal, S., and Beyler, C. 1990. Correlating Pre-flashover Room Fire Temperatures. *Journal of Fire Protection Engineering* 2(2): 33-48.

Dua, D. and Graff, C. 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Dunn, V. 2015. Fire Engineering Books: Safety and Survival on the Fireground. Pennwell Books. United States.

Fahy, R. F., Petrillo, J. T., and Molis, J. L. 2020. Firefighter Fatalities in the United States - 2019, Report, National Fire Protection Association, Quincy, MA.

Garrido, M.A. and Font, R. 2015. Pyrolysis and combustion study of flexible polyurethane foam. *Journal of Analytical and Applied pyrolysis* 113: 202-215.

Guo, L., Li, N., Jia, F., Lei, Y., and Lin, J. 2017. A Recurrent Neural Network based Health Indicator for Remaining Useful Life Prediction of Bearings. *Neurocomputing* 240: 98-109.

Graves, A., and Schmidhuber, J. 2005. Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures. *Neural networks* 18(5-6): 602-610.

Hamins, A.P., Bryner, N.P., Jones, A.W., and Koepke, G.H. 2015. Research roadmap for smart fire fighting. Special Publication 1191. National Institute of Standards and Technology, Gaithersburg, MD.

Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural computation* 9(8): 1735-1780.

Kim, H.J. and Lilley, D.G., 2002. Heat release rates of burning items in fires. *Journal of propulsion and power* 18(4): 866-870.

Kurzawski, A.J., and Ezekoye, O.A. 2020. Inversion for Fire Heat-Release Rate Using Heat Flux Measurements. *Journal of Heat Transfer* 142(5).

Lattimer, B.Y., Hodges, J.L., and Lattimer, A.M. 2020. Using machine learning in physics-based simulation of fire. *Fire Safety Journal*: 102991.

Li, N., Lee, E.W., Cheung, S.C., and Tu, J. 2019. Multi-fidelity Surrogate Algorithm for Fire Origin Determination in Compartment Fires. *Engineering with Computers*: 1-18.

Liu, Y., and Wu, Y.F.B.,2018. Early Detection of Fake News on Social Media through Propagation Path Classification with Recurrent and Convolutional Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Madrzykowski, D. and Weinschenk, C. 2019. Impact of fixed ventilation on fire damage Patterns in full-scale structures. Report. Underwriters Laboratories Inc. Columbia, MD.

Mattern, J. L. 2017. It’s Official: Americans Are Obsessed with Ranch Homes. https://www.countryliving.com/real-estate/a44039/

ranch-home-united-states/, 2017. Accessed: 2020-08-06.

McCaffrey, B.J., Quintiere, J.G., and Harkleroad, M.F. 1981. Estimating Room Temperatures and the Likelihood of Flashover using Fire Test Data Correlations. *Fire Technology* 17(2): 98-119.

McGrattan, K. 2007. Verification and Validation of Selected Fire Models for Nuclear Power Plant Applications (7): Fire Dynamics Simulator (FDS). Final Report. National Institute of Standards and Technology, Gaithersburg, MD.

McKinnon, M., Weinschenk, C. and Madrzykowski, D. 2020. Modeling Gas Burner Fires in Ranch and Colonial Style Structures. Report. Underwriters Laboratories Inc. Columbia, MD.

National Fire Protection Association, 2002. NFPA 72–National Fire Alarm Code, 2002 Edition. National Fire Protection Association, Quincy, MA.

Overholt, K. J., and Ezekoye, O. A. 2012. Characterizing Heat Release Rates using an Inverse Fire Modeling Technique. *Fire Technology* 48(4):893-909.

Peacock, R.D., Reneke, P.A., Bukowski, R.W., and Babrauskas, V. 1999. Defining Flashover for Fire Hazard Calculations. *Fire Safety Journal* 32(4):331-345.

Peacock, R.D., McGrattan, K.B., Forney, G.P. and Reneke, P.A. 2015. CFAST–Consolidated Fire And Smoke Transport (Version 7) Volume 1: Technical Reference Guide. Technical Note 1889, National Institute of Standards and Technology, Gaithersburg, MD.

Ramesh, N. and Venkateshan, S.P. 1999. Effect of surface radiation on natural convection in a square enclosure. *Journal of thermophysics and heat transfer* 13(3): 299-301.

Reneke, P.A., Bruns, M., Gilbert, S.W., MacLaren, C.P., Peacock, R.D., Cleary, T.G. and Butry, D.T. 2019. Towards a Process to Quantify the Hazard of Fire Protection Design Alternatives. NIST TN-2041. National Institute of Standards and Technology, Gaithersburg, MD.

Richards, R., Munk, B., and Plumb, O. 1997. Fire Detection, Location and Heat Release Rate Through Inverse Problem Solution. Part I: theory. *Fire Safety Journal* 28(4): 323–350.

Stowell, F. M. and Murnane, L. 2017. Essentials of Fire Fighting and Fire Department Operation. 7th Edition. International Fire Service Training Association.

Tam, W.C., Fu, E.Y., Peacock, R., Reneke, P., Wang, J., Li, J. and Cleary, T. 2020. Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard. *Fire Technology*: 1-22.

Thomas, P.H., Bullen, M.L., Quintiere, J.G., and McCaffrey, B.J. 1980. Flashover and Instabilities in Fire Behavior. *Combustion and Flame* 38:159-171.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*: 5998-6008.

Walton, W.D., Thomas, P.H., and Ohmiya, Y. 2016. Estimating temperatures in compartment fires. In SFPE handbook of fire protection engineering: 996-1023. Springer, New York.

Wu, X., Park, Y., Li, A., Huang, X., Xiao, F., and Usmani, A. 2020. Smart Detection of Fire Source in Tunnel Based on the Numerical Database and Artificial Intelligence. *Fire Technology*.

Yu, C., Wang, S., Lin, C., Chou, K., Lai, C., and Chen, T. 2012. Fire zone/Field Model Performance Based Investigation in Fire Flashover Phenomenon. In *2nd International Conference on Consumer Electronics, Communications and Networks*: 3258-3261.

Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., and Morency, L.P. 2018. Multi-Attention Recurrent Network for Human Communication Comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Zhang, C., Grosshandler, W., Sauca, A., and Choe, L., 2020. Design of an ASTM E119 Fire Environment in a Large Compartment. *Fire Technology* 56(3): 1155-1177.

Zhang, G., Zhu, G., Yuan, G., and Huang, L. 2014. Methods for Prediction of Temperature Distribution in Flashover Caused by Backdraft Fire. *Mathematical Problems in Engineering*.

Tam, Andy; Fu, Eugene Yujun; Peacock, Richard D.; Reneke, Paul A.; Wang, Jun; Ngai, Grace; Leong, Hong Va; Cleary, Thomas. ”Predicting Flashover Occurrence using Surrogate Temperature Data.” Presented at 35th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI-21). February 02, 2021 - February 09, 2021.

PERFORMANCE EVALUATION OF X-RAY COMPUTED TOMOGRAPHY INSTRUMENTS: SENSITIVITY TO DETECTOR AND STAGE ERRORS-TRENDS FOR DIFFERENT MAGNIFICATIONS

Prashanth Jaganmohan[1,2], Bala Muralikrishnan[1], Meghan Shilling[1] and Edward P. Morse[2]
[1]Sensor Science Division
National Institute of Standards and Technology
Gaithersburg, MD, USA
[2]Center for Precision Metrology, Department of Mechanical Engineering
University of North Carolina at Charlotte
Charlotte, NC, USA

INTRODUCTION

In recent times, industrial metrology has been experiencing a steady increase in the use of X-Ray Computed Tomography (XCT) as a means for inspection, primarily when dealing with complex parts that would be time consuming to measure by traditional inspection procedures, or parts with internal features that simply cannot be accessed or measured by contact-based methods. This increasing demand brings forth the need for development of standardized test procedures to evaluate the performance of XCT instruments and support claims of metrological traceability. To meet these needs, the International Organization for Standardization (ISO), the American Society of Mechanical Engineers (ASME), and ASTM International have been working independently to develop performance evaluation standards for XCT systems. Current drafts of these standards recommend testing at different measurement volumes (i.e., at different stage and detector positions), but do not provide specific guidance as to where these measurement volumes are to be located or all of the test positions within these volumes [1-2]. Part of the reason for the lack of such specific information could be the cumbersome task of testing all conditions experimentally or through extensive radiographic simulations. The work described in this paper overcomes the difficulty of such an unrealistic workload of simulations since a significantly faster simulation method is used.

One of the main goals associated with the development of performance evaluation standards is to design test procedures that are sensitive to all or as many known error sources as possible. Thus, it becomes imperative to identify all significant error sources, understand their influence on dimensional measurements and accordingly make recommendations on methods to capture them. Several error sources in XCT systems have been identified and discussed in the VDI/VDE 2630-1.2 [3]. The National Institute of Standards and Technology (NIST) has conducted studies [4-5] on the effect of one particular type of error source, namely uncorrected geometry errors in cone-beam XCT systems. In that work, the sensitivities of sphere center-to-center distance errors and sphere form errors to various geometrical error sources were explored for one position of the detector and the rotation stage. Based on results obtained by a NIST-developed rapid simulation method, i.e., the single point ray tracing method (which was validated against more traditional radiograph-based reconstruction method), sphere center-to-center distance error sensitivities and form error sensitivities of spheres were reported.

The present work is an extension of the research in [4-5] by repeating the simulations at several combinations of the source-stage (also known as source-object) and source-detector distances. For a given error source under consideration, the combination of these distances at which the highest sensitivities can be obtained, is first identified. Such a combination of distances establishes the geometric magnification, given by the ratio of the source-detector distance to the source-stage distance. This combination also establishes the dimensions of the measurement volume for a given XCT instrument. Within such a measurement volume, the measurement line (i.e., the position and orientation of a center-to-center distance segment in the measurement

Jaganmohan, Prashanth; Muralikrishnan, Bala; Shilling, Meghan; Morse, Ed. "Performance Evaluation of X-ray Computed Tomography Instruments: Sensitivity to Detector and Stage Errors - Trends for Different Magnifications." Presented at 35th Annual Meeting of the ASPE. October 20, 2020 - October 22, 2020.

volume) that produces the maximum center-to-center distance error (and therefore, the highest sensitivity) for a given magnitude of the considered error source is identified. This process is repeated for all error sources studied here. Therefore, for every error source, the source-stage distance, source-detector distance, and the line of highest sensitivity in the corresponding measurement volume are determined here.

SIMULATION SETUP

All simulations in the present work are carried out in MATLAB (R2018b)[1]. The simulated reference object consists of a set of 125 spheres distributed into five horizontal planes. Each plane consists of one centrally located sphere, 16 spheres arranged in an outer circle, and eight spheres arranged in an inner circle whose radius is half the radius of the outer circle, as shown in Figure 1. This is the same arrangement described in [4]. Note that in Figure 1, the positions of the source and detector are to scale, but the size of the detector is not. The coordinate system is centered at the source, with axis directions as shown in Figure 1. Detailed descriptions on how these axes are defined to establish the coordinate system can be found in [4]. While the object was of a single fixed size in the study conducted earlier [4-5], here, the diameter and height of the overall cylindrical shape are a function of the source-stage and source-detector distances. That is, the simulated object is scaled for each combination of source-stage and source-detector distance pair so as to fill 98% of the area of a 250 mm x 250 mm detector. Such scaling is done to obtain the largest magnitudes of distance errors possible for each combination of stage and detector positions. The diameters of the spheres are also scaled accordingly (varying from 3.40 mm at the smallest measurement volume to 18.68 mm at the largest measurement volume) while making sure their projected images still fit on the detector.

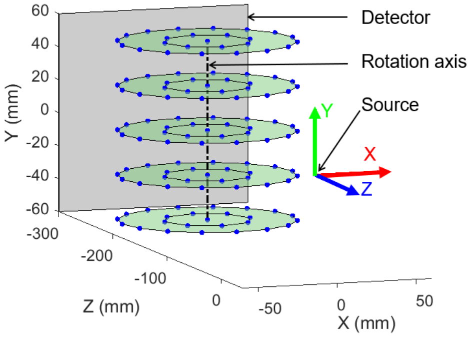


FIGURE 1. Reference object containing 125 spheres chosen for simulation.

The geometric error sources associated with the detector include errors in location along the three mutually perpendicular axes, and errors in angular position about the same axes. These are shown in Figure 2. Errors associated with the stage include a location error along the Z axis (connecting the source and rotation axis), and error motions of the stage, which includes encoder scale, axial, radial, and wobble errors. All the error motions of the stage are assumed to have harmonic components and therefore are represented as sine and cosine functions of the rotation stage indexing angle. Harmonics of orders one through ten are included in this study. The error sources associated with the stage are shown in Figure 3. Detailed descriptions of these error sources and the coordinate system in which these error souces are defined, can be found in Ferrucci et al. [6-7] and NIST studies [4-5].

METHODOLOGY

The methodology used for the simulations consists of the single point ray tracing method introduced in [4]. This method has been proven to be a faster and more practical alternative to the full X-ray CT tomographic reconstruction methods for estimating the effects of geometric errors using sphere-based artifacts. In the single point ray tracing method, only the centers of the spheres are projected on to the detector using a known set of parameters describing the geometric errors of the system. A locus representing the motion of each projected center on the detector is obtained by determining the projection for each angular position of the stage until a full revolution is complete. The loci obtained are used in a back-projection algorithm to estimate the location of each sphere through a least-squares minimization. The reconstructed

[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.

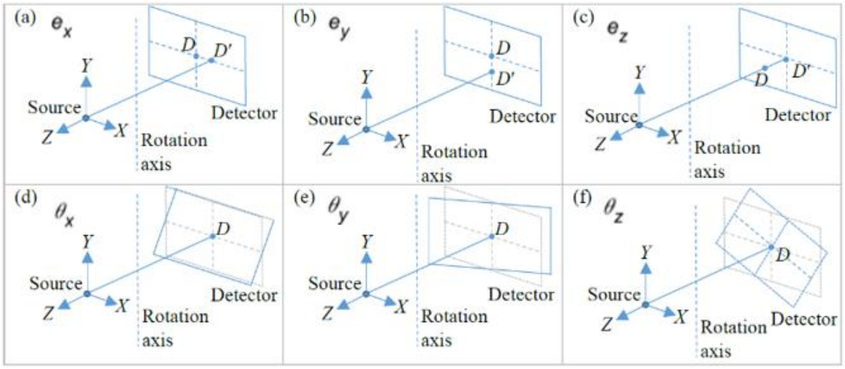


FIGURE 2. Geometrical error sources associated with the detector. (a) Detector location error parallel to X axis. (b) Detector location error parallel to Y axis. (c) Detector location error along the Z axis. (d) Detector rotation error about an axis parallel to X axis. (e) Detector rotation error about an axis parallel to Y axis. (f) Detector rotation error about the Z axis.

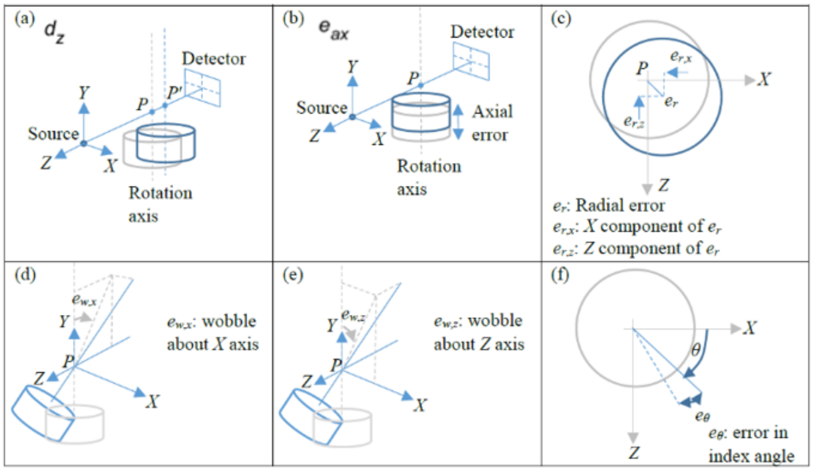


FIGURE 3. Geometrical error sources associated with the stage. (a) Error in the location of the rotation axis along Z. (b) Axial error motion. (c) Radial error motion components along X and Z. (d) The X component of the wobble error. (e) The Z component of the wobble error. (f) Error in the angular position of the stage.

sphere centers obtained in this way allow the calculation of the center-to-center distances for each pair of the 125 sphere centers. Circles comprising discrete points are drawn normal to each ray from the source to the detector with their centers located on the previously identified least-squares centers. The points lying on the outer surface of the resulting point cloud are considered for form error calculation.

When analyzing the effect of a particular error source, its actual value is fed into the forward projection, for example, the actual pitch angle error of the detector is considered in order to generate the simulated sphere centers for each angular position. However, the back-projection, or reconstruction algorithm, assumes an ideal geometry, i.e., the absence of that geometric error. In this example, this would mean the ideal pose of the detector, i.e., zero pitch. This discrepancy between actual and assumed parameters defines the magnitude of simulated geometry errors and results in sphere center-to-center distance errors and sphere form errors. In this way, the effect of all geometry errors including detector errors and stage errors on the center-to-center distance errors and form errors of the spheres on the reference object are studied. To understand the effect of magnification, such simulation studies are performed at several positions of the stage and detector throughout the working volume.

For each error source under consideration, the pair of spheres that produced the highest center-to-center distance error is identified for the combination of source-stage and source-detector distances identified in [4]. The line joining the previously identified pair of spheres constitutes the line of highest sensitivity, i.e. center-to-center distance error (in mm) per mm or degree of geometric error, and is tracked across all combinations of the stage and detector positions.

The stage is positioned at varying distances (*d*) from the source, starting from 200 mm and increasing in steps of 100 mm up to a maximum of 1100 mm. For each chosen position of the stage, the source-detector distance (*D*) is varied from *d*+100 mm to 1200 mm, in steps of 100 mm.

RESULTS AND DISCUSSION

For all the error sources under consideration, the highest distance error sensitivity is observed to occur at one of two configurations. In the first configuration, the detector is positioned as close to the source as possible, and the stage is then positioned as close to the detector as possible. For the simulations conducted, this corresponds to a source-stage distance (*d*) of 200 mm and a source-detector distance (*D*) of 300 mm.

The second configuration that tends to produce the highest distance sensitivity is where the stage is positioned as far from the source as possible and the detector is then positioned as close to the stage as possible. Here, this configuration corresponds to *d*=1100 mm and *D*=1200 mm. In traditional XCT systems, this second configuration usually corresponds to the largest measurement volume possible for a given instrument.

The solid lines shown in Figure 4 indicate the lines that produced that highest distance error sensitivity for the six detector errors and the Z location error of the rotation axis. In cases where multiple lines are shown for a single error source, it simply indicates that all the highlighted lines are equivalent in capturing the highest sensitivity shown. The values of *d* and *D* at which this highest sensitivity is observed are also mentioned in each case.

Figures 5 through 8 show similar illustrations of sensitive lines for the error motions of the stage. These errors are represented by sine and cosine components of orders 1 to 10. However, the magnitudes of sensitivities corresponding to fifth order and higher harmonics are found to be negligible. Therefore, the sensitive lines for the first four orders are shown here. Further, only the cosine components of the error sources are shown. The sensitive lines for the sine components are observed to have similar orientations but rotated by 90 degrees about the rotation axis.

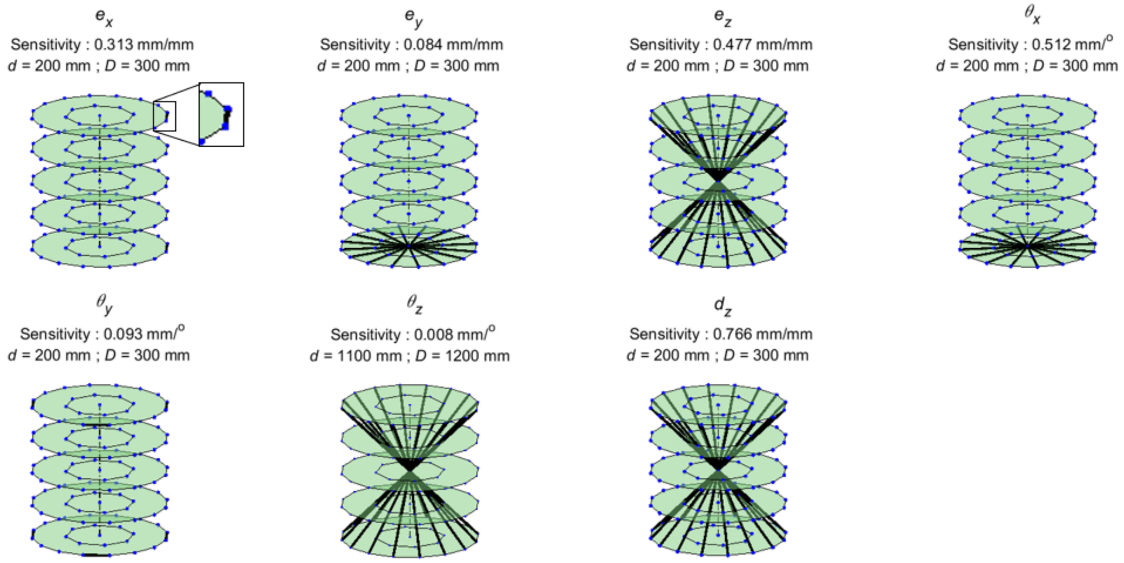

e_x
Sensitivity : 0.313 mm/mm
d = 200 mm ; D = 300 mm
e_y
Sensitivity : 0.084 mm/mm
d = 200 mm ; D = 300 mm
e_z
Sensitivity : 0.477 mm/mm
d = 200 mm ; D = 300 mm
θ_x
Sensitivity : 0.512 mm/°
d = 200 mm ; D = 300 mm
θ_y
Sensitivity : 0.093 mm/°
d = 200 mm ; D = 300 mm
θ_z
Sensitivity : 0.008 mm/°
d = 1100 mm ; D = 1200 mm
d_z
Sensitivity : 0.766 mm/mm
d = 200 mm ; D = 300 mm

FIGURE 4. Sensitive lines for detector errors and Z location error of rotation axis.

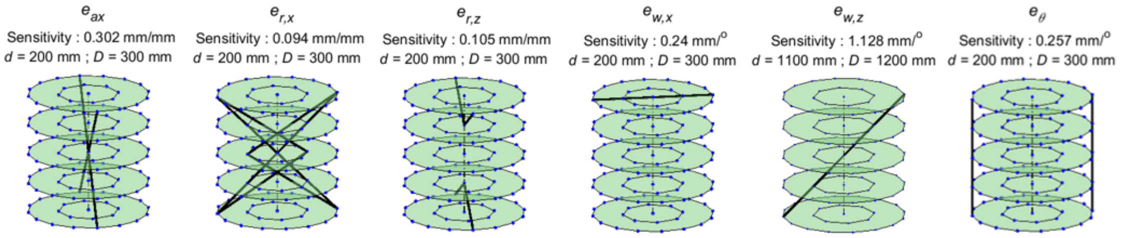


FIGURE 5. Sensitive lines for first order cosine components of stage error motions

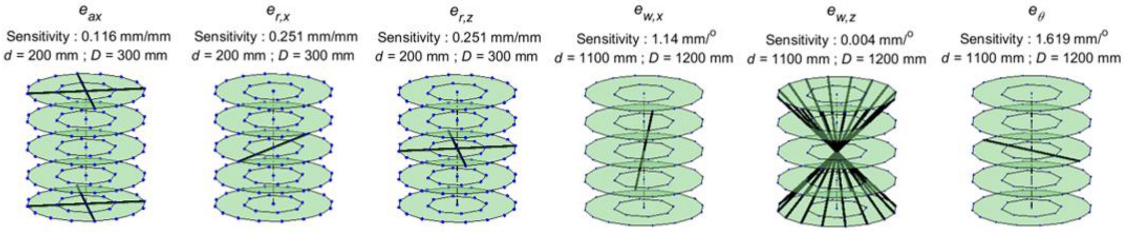


FIGURE 6. Sensitive lines for second order cosine components of stage error motions

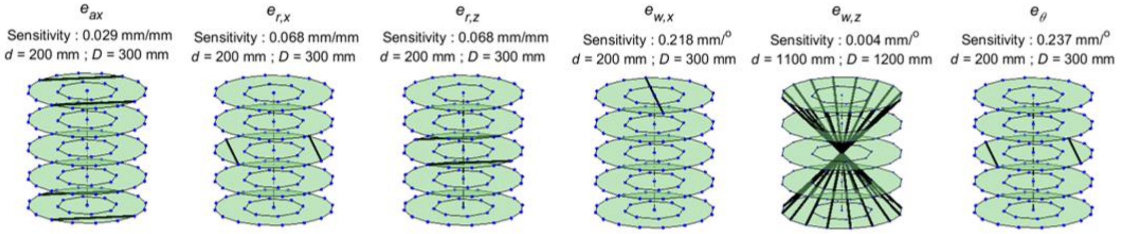


FIGURE 7. Sensitive lines for third order cosine components of stage error motions

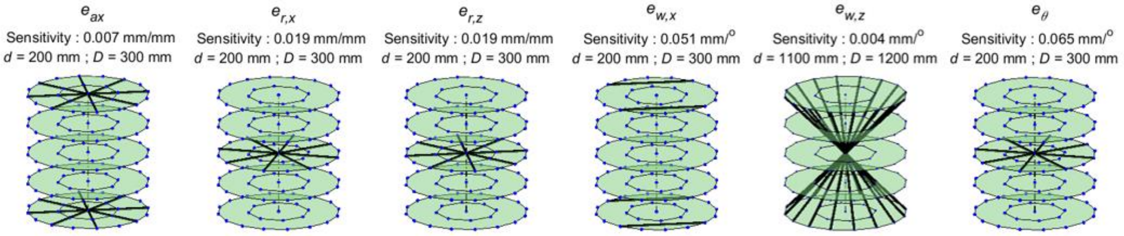


FIGURE 8. Sensitive lines for fourth order cosine components of stage error motions

CONCLUSIONS

Simulations have been conducted on a set of 125 spheres distributed throughout the measurement volume. Such simulations have been repeated for various positions of the stage and detector. The pair of spheres that produced the largest error per unit magnitude of a given geometric error source have been identified. This pair defines the line of highest sensitivity for that error source, i.e., this line defines the position and orientation in the measurement volume that is most sensitive to the given error source. Some key takeaways are:

- All error sources have their maximum center-to-center distance error sensitivity occurring at one of two configurations, one with the detector as close as possible to source and stage as close as possible to detector, and the other with the stage as far as possible from the source and the detector as close as possible to the stage. Notice that in both cases, i.e., whether the stage is as close to the source or the stage as far away from the source, the magnification is the lowest possible.

- Most of the sensitive lines identified are body diagonals, vertical lines and diametrical lines at various horizontal levels, but it is interesting to note that some other ‘uncommon’ lines were also identified- such as some horizontal lines that are not diametrical or otherwise intuitive.

Future work could include extension of such a study to other error sources not included here. Further, studies can be conducted on designing suitable artifacts and making recommendations on a minimum set of test positions and measurement lines by which all known error sources can be captured to a satisfactory degree. In addition, experimental validation of these results by measurements on a physical artifact are also envisioned.

REFERENCES

[1] Draft ISO/CD 10360-11 - Geometrical product specifications (GPS) — Acceptance and reverification tests for coordinate measuring machines (CMM) — Part 11: CMMs using the principle of computed tomography (CT)

[2] Draft ASME B89.4.23 - X-Ray Computed Tomography (CT) Performance Evaluation Standard.

[3] VDI/VDE 2630 -1.2: Computed tomography in dimensional measurement; Influencing variables on measurement results and recommendations for computed tomography dimensional measurements.

[4] Muralikrishnan B, Shilling M, Phillips S, Ren W, Lee V, Kim F. X-ray Computed Tomography Instrument Performance Evaluation, Part I: Sensitivity to Detector Geometry Errors. Journal of Research of the National Institute of Standards and Technology. 2019; 124:124014 1-16.

[5] Muralikrishnan B, Shilling M, Phillips S, Ren W, Lee V, Kim F. X-ray Computed Tomography Instrument Performance Evaluation, Part II: Sensitivity to Rotation Stage Errors. Journal of Research of the National Institute of Standards and Technology. 2019; 124:124015 1-13.

[6] Ferrucci M, Ametova E, Probst G, Craeghs T, Dewulf W. Sensitivity of CT dimensional measurements to rotation stage errors. Proceedings of the 8th Conference on Industrial Computed Tomography (Wels, Austria), 2018.

[7] Ferrucci M, Ametova E, Carmignato S, Dewulf W. Evaluating the effects of detector angular misalignments on simulated computed tomography data. Precision Engineering, 2016; 45 230-241.

A Machine Learning Based Scheme for Dynamic Spectrum Access

Anirudha Sahoo
Communications Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, Maryland, USA
anirudha.sahoo@nist.gov

***Abstract*—In this paper, we present a machine learning (ML) based dynamic spectrum access (DSA) scheme which can be used in a system in which the primary user (PU) spectrum occupancy can be represented as a sequence of busy (on) and idle (off) periods. We use real world data collected from Long Term Evolution (LTE) systems at two locations for our study. We experiment with different feed forward artificial neural network (ANN) architectures to choose from for our DSA scheme. A simple perceptron based ANN architecture was determined to provide good performance. We compare performance of our ML based DSA scheme with a traditional DSA scheme based on analytical model that uses survival analysis. Our results show that our ML based scheme outperforms the survival analysis based scheme in terms of utilization of idle periods. In terms of probability of interference to the PU, our scheme is better in some configurations and slightly worse in some other configurations.**

***Index Terms*—Dynamic spectrum access, spectrum sharing, machine learning, artificial neural network.**

I. Introduction

There is an acute scarcity of spectrum, especially below 6 GHz, due to static spectrum allocation policy. However, the spectrum utilization in some of the bands is low [1]. Thus, there is scope for increasing spectrum utilization in those bands through sharing. Hence, dynamic sprectrum access (DSA) has been proposed as a way of sharing spectrum with the incumbents to improve spectrum utilization. An incumbent is usually referred to as primary user (PU) and the user that shares the spectrum using DSA is referred to as secondary user (SU). Although there are different DSA schemes, in this study we focus on DSA based on opportunistic spectrum access (OSA). In an OSA based DSA system, an SU opportunistically transmits when the spectrum is idle due to inactivity of the PU. But the SU should vacate the spectrum before the PU reappears to avoid interfering with the PU. Since typically there is no communication between PU and SU, the SU has to predict when the PU might reappear, i.e., when the idle period might end. Based on this prediction, the SU decides how long it should transmit or if a request to transmit for a certain duration be granted. Thus, the performance of such a DSA system depends on how well the SU predicts the spectrum occupancy.

Machine learning (ML) has been used for various kinds of prediction problems, e.g., image recognition, natural language processing. So, it is only natural that ML-based techniques are used for DSA. Many DSA schemes have been designed based on analytical models. But analytical models have limitations. Normally, analytical models make certain assumptions to fit the DSA scenario to the model. These assumptions may not hold for certain datasets or for certain scenarios. Sometimes, there may be interdependencies in the input parameters which may not be adequately captured by the analytical models. These limitations can lead to performance degradation of the analytical models. ML based models, on the other hand, can overcome these limitations by having appropriate architecture to accommodate interdependencies of input parameters and by having enough training data in different scenarios to produce good results in many scenarios. Thus, in general, ML based models, with appropriate architecture, input features and training dataset can produce better results in many different scenarios. Although we found some ML-bsed schemes for DSA in the literature (see Section II), to the best of our knowledge, a very simple feed forward artificial neural network (ANN) based approach for DSA is missing. Hence, in this work, we devise a DSA scheme using basic feed forward ANN which can be used in a system in which spectrum occupancy of the PUs can be represented as a sequence of busy (on) and idle (off) periods. We used a Long Term Evolution (LTE) network as the PU system and considered the spectrum (channels) used in the uplink for DSA. For our study, we used real world LTE uplink data collected at two locations and devised an ANN based DSA scheme on a given channel of the LTE uplink. A channel in LTE is the smallest allocable range of frequency which is 180 kHz. We experimented with different input features for the ANN and decided on the set of input features that provided good performance. We also experimented with a few different ANN architectures and found that performance of a simple perceptron based ANN is comparable to that of more complex deep neural network (DNN) architectures. Hence, we chose a simple perceptron based ANN for our DSA scheme. The results from our experiments show that our ANN based scheme is able to utilize PU idle periods (whitespace) quite well with low probability of interference to the PU. We compared the performance of our ANN based scheme with a traditional DSA scheme based on an analytical model. The traditional DSA scheme, which is based on survival analysis, was propsed in [2]. From the experimental results, we observed that utilization of idle periods using our ANN based scheme was always higher than the survival ananlysis based scheme in all configrations. However, the probability of interference was lower for some configurations and slightly higher for some

Sahoo, Anirudha. "A Machine Learning Based Scheme for Dynamic Spectrum Access." Presented at 2021 IEEE Wireless Communications and Networking Conference (WCNC 2021), Nanjing, CN. March 29, 2021 - April 01, 2021.

other configurations. Therefore, this study shows that it is possible to design an effective DSA system using a simple feed forward ANN.

II. Related Work

There have been quite a bit of work reported in the literature on prediction of spectrum occupancy. A Partially Observable Markov Decision Process (POMDP) to predict spectrum occupancy has been proposed in [3]. The DSA scheme proposed in [4] uses expected remaining off time for prediction. Some schemes indirectly predict the spectrum occupancy by limiting the transmission duration of SUs based on some constraint. In [5], maximum bound on probability of interference to PU is used as a constraint to compute the duration of transmission of an SU. Spectrum occupancy is modelled as an alternating renewal process in [6] and residual idle time of the idle duration (in which SU request arrives) is used to indirectly predict when the spectrum will be busy again. Multiarm Bandit model has also been used in OSA channel access [7], [8]. Pattern mining of occupancy data has been used to predict channel idle period [9], [10]. DSA algorithms based on survival analysis have been presented in [2]. While [2] proposes channel access algorithms in time domain only, the survival analysis based algorithms presented in [11] are meant for DSA in time and frequency (or channel) domains.

Machine learning (ML) based techniques have been shown to be very effective with prediction in various fields (e.g., image recognition, natural language processing). It is no surprise that researchers have used ML to design DSA schemes. A Q-learning based algorithm to improve DSA performance in terms of channel throughput has been proposed in [12]. Faganello et al. proposed improvements to Q-learning algorithm to design three schemes for DSA in a dynamic industrial cognitive radio network [13]. Deep Q-learning was proposed as an ML approach by combining Q-learning and neural networks [14]. Such an architecture is called deep Q-learning network (DQN). In the DQN based algorithm proposed in [15], an SU avoids heavy interference regions and selects efficient frequency hopping pattern to dynamically access spectrum. In [16], a deep recurrent neural network is used to learn the time varying distribution of user traffic in a Land Mobile Radio (LMR) network, which is then used to determine best spectrum assignment and sharing strategies. In the graph neural network based approach for DSA in femtocells proposed in [17], the graph neural network maps the traffic load to a channel access scheme. In this work, a multiagent reinforcement learning framework is used for training and to predict channel quality.

III. Problem Formulation

In an opportunistic dynamic spectrum access (DSA) system, the spectrum occupancy by the PU can be represented as a series of *busy* and *idle* periods. An SU accesses the spectrum when it is idle and should finish transmission before the spectrum is occupied by the PU, i.e., before the next busy period starts. Since, it is not known exactly when the next busy period will start, the SU has to rely on some kind of prediction. Based on the accuracy of the prediction method, sometimes the SU may not finish transmission before the next busy period, which leads to interference with the PU. An efficient opportunistic DSA system keeps this interference to a low value.

Figure 1 shows an example scenario of opportunistic DSA. An SU request to transmit arrives at time instant 'A'. If the request is to transmit for duration τ_1 and the prediction system grants the request, then there will be no interference, i.e., the transmission will be successful. In this case the prediction system made the correct decision. On the other hand, if the requrest is to transmit for duration τ_2 and the prediction system grants the request, then there will be interference. In this case, the prediction system made a wrong decision; the correct decision is to deny the request. Note that in the first case, the prediction system could make a wrong decision and deny the request for transmission of duration τ_1, which would be a lost opportunity.

ML has been used in prediction systems [18], [19], [20]. Supervised learning using ANN is a good candidate for prediction. In such systems, an ANN is trained with a training dataset. The training dataset contains ground truth about the prediction for a set of inputs. For our problem, the training dataset contains some inputs for which the prediction is correct and some other inputs for which the prediction is wrong. Once the ANN is trained, it is used for prediction with input data which it has not seen before. However, for our application, we need to decide what should be the input features for the ANN. We also need to choose an ANN architecture which provides good performance. We make these choices by running experiments against different input featues and different ANN architectures.

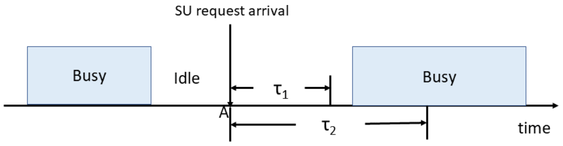


Fig. 1: Opportunistic DSA Example

IV. Evaluation

Our goal is to design a prediction system based on ANN that can facilitate opportunistic DSA for an SU. The ANN is trained using a training dataset. The training dataset consists of a set of inputs and the corresponding known output. In our case, the output is binary indicating whether the the SU request was granted or denied. Thus, our ANN based prediction system is a supervised binary classification system. Once the ANN is trained with the training dataset, when a request to transmit for duration τ arrives at an SU, the SU presents the input corresponding to the request to the ANN and gets an output. An output of 0 implies denying the request, whereas an output of 1 means the request is granted.

Sahoo, Anirudha. "A Machine Learning Based Scheme for Dynamic Spectrum Access." Presented at 2021 IEEE Wireless Communications and Networking Conference (WCNC 2021), Nanjing, CN. March 29, 2021 - April 01, 2021.

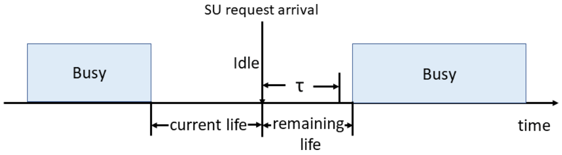


Fig. 2: Illustration of Some Input Features

A. Input Feature Selection

Selecting approporiate input features for an ANN is very important for its performance. In our case, given that the SU request arrives during an idle period, we know the *current life* of the idle period at the time of the request arrival. Current life of an idle period is the duration from the beginning of the idle period to the current time. Note that if we knew the *remaining life* of the idle period (the duration from current time until the end of the idle period), then we could deterministically decide whether to grant or deny the request which would eliminate the need for a prediction system. Hence, remaining life cannot be an input to the ANN. Figure 2 illustrates current life and remaining life of an idle period. Current life is undoubtedly an important feature for our ANN. Previous idle periods may also have correlation with the current idle period. Hence, we chose few previous idle periods as input features. Through experiments we found that ten previous idle periods is a good number. Lastly, the duration of SU request (τ) is also an important parameter for prediction, So, we chose current life of idle period, previous ten idle periods and the duration of SU request as our input features for the ANN. Thus, there were twelve nodes in the input layer.

B. Feature Scaling

We noticed that the values of the input features were quite varied, i.e., while some were small in value, others were very large. Hence, the values of input features were normalized to a value between 1 and 100.

C. Data Collection

For our experiements, we collected real world LTE data at two locations.

First set of data was collected indoors inside one of the labs at the National Institute of Standards and Technology (NIST). An Ettus Universal Sofware Radio Peripheral (USRP) [1] running USRP hardware driver (UHD) version 003.009.001 was fitted with a small 10.78 cm rubber duck antenna. Using GNU Radio version 3.7.9rc1, complex-valued I/Q samples were collected every 80 ns with a sampling rate of 12.5 MHz in the Band 17 with a 10 MHz uplink (UL) LTE band at center frequency 709 MHz. In every 50 μs period, 625 consecutive I/Q samples were collected and power spectrum over that period was computed. Average power spectra over 20 consecutive periods (equal to 1 ms) were then computed. 56 point power spectrum for each 1 ms period was then generated by binning the coefficients. Each power spectrum coefficient was rounded to the nearest integer and represented as an 8 bit integer. Each of these coefficients represents power (in dB) over a 180 kHz frequency range. 50 LTE channels in a 10 MHz UL are represented by the middle 50 coefficients. These power values were then converted to binary on or off (0 or 1) by applying a noise threshold to each of the 50 channels. The noise threshold was determined as follows. Samples were collected for a one hour period from the USRP after connecting a matched-load terminator to the receiver port. The noise threshold was determined to be the level at which 1 % of the sample values were above the threshold. This corresponded to probability of false alarm (PFA) of 1 % [21]. We collected the data on two week days at the same time of the day. The idea is to use one dataset as training data and the other as test data. The first dataset was collected on a Monday from 3 PM to 4 PM local time, whereas the second dataset was collected on the nextday (Tuesday) from 3 PM to 4 PM. These timings were chosen assuming that there would be high LTE traffic during those times.

The second dataset was collected in the Philadelphia metropolitan area near University of Pennsylvania (we will refer to this as *UPENN* data). We requested the administrator of the CityScape spectrum [22] to collect LTE uplink spectrum data. LTE uplink data was collected on a weekday at 1:17 PM local time for an hour. Additional processing of the I/Q samples was carried out to bring the data to the same format as the data collected at NIST. Finally, the CityScape datasets was converted to binary occupancy sequences using the noise threshold recommended by the CityScape administrator. The noise threshold at the CityScape site was determined by measuring the power level when no UE was believed to be communicating. The one hour worth of data is split into two equal parts of half an hour each. One part was used as training data whereas the other was used as test data.

D. Generation of Training and Testing Dataset

Given the spectrum occupancy data, the training dataset was created as follows. SU requests were simulated to arrive with a Poisson distribution, i.e., the interarrival times of the requests were exponentially distributed with a given mean arrival rate. Experiments were repeated with different mean arrival rate. For a given SU request, the requested transmission time was uniform randomly chosen between 1 ms and 10 ms. Since the arrival time of the SU request was known, current life of the idle period was computed and then the remaining life (see Figure 2) of the current idle period was computed. The last ten idle periods, current life of this idle period and the SU request duration form an input set. If this remaining idle duration was greater than the SU's requested transmission time, then the transmission was successful and the output label for this input set was set to 1, otherwise it was set to 0. The training data was randomly shuffled and then split into two equal halves. The first half was used to train the ANN and the second half was used as validation data.

[1] The identification of any commercial product or trade name does not imply endorsement or recommendation 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 testing dataset was built exactly with the same manner, but on a different collected dataset. For example, for datasets collected at NIST, dataset of one day was used to generate training data whereas dataset of the other day was used to generate testing data. Similarly for the UPENN dataset, a half hour dataset was used to generate training data whereas the other half hour dataset was used to generate test data.

E. Choosing an Appropriate ANN Architecture

We need to decide what kind of ANN architecture and configuration are appropriate for our application. Towards that goal, we used the binary occupancy data of channel 5 of the NIST dataset. The first day's data was used for training and the second day's data was used for testing. Table I lists the different architectures and the corresponding results. We used keras on a server running linux to implement our code. For all architectures, the training was stopped after 100 epochs. Reducing the number of neurons from one hidden layer to the next (as we move towards the output layer) has been reported to be a good rule of thumb for ANN architecture [23]. Hence, we started with a Deep Neural Network based ANN architecture with three hidden layers, the first with 20 nodes, the second with 10 and third with 5 nodes. The training accuracy with this architecture was 78.47 %, whereas validtion accuracy was 77.57 %. Testing accuracy, in this case, was 91.8 %. We then simplified the architecture by having only two hidden layers. In this case, the training and validation accuracy came down only a little, but testing accuracy increased to 95.9 %. So, we continued to simplify the architecture as shown in the table. When we reduced the number of hidden layers and also reduced the number of nodes in the hidden layer, we did not see any difference in the testing accuracy, whereas the training and validation accuracy decreased slightly. So, we went all the way down to no hidden layer, i.e., just a single perceptron. Even with a single perceptron, the training and validation accuracy went down by a very small amount, whereas the testing accuracy remained the same. Hence, a single perceptron based architecture was deemed appropriate for our application. So, performance evaluation of our scheme was done with a single perceptron architecture as shown in Figure 3. Once trained, this ANN architecture can predict whether an SU request should be accepted or not in O(N) time, where N is the number of inputs to the perceptron. Therefore, run time of our DSA scheme is low with just twelve inputs.

F. Notations Used

As mentioned earlier, for the NIST data, data collected on one day was used for training whereas the data collected on the other day was used for testing. When NIST data is used in our experiments, we denote a configuration as NIST_*train*_*test*, where *train* and *test* are either *day1* or *day2* depending on which day's data is used for training and which day's data is used for testing. For example, configuration *NIST_day1_day2* implies day1 data is used for training the perceptron and day2 data is used for testing. Similarly, for data collected at UPENN, the configuration is denoted as UPENN_*train*_*test*, where *train* and *test* are either *hh1* or *hh2*, depending on whether first half hour (hh) or second half hour data is used for training. For example, configuration *UPENN_hh2_hh1* represents UPENN data collected in the seocnd half hour used as training data, whereas data collected in the first half hour used as testing data.

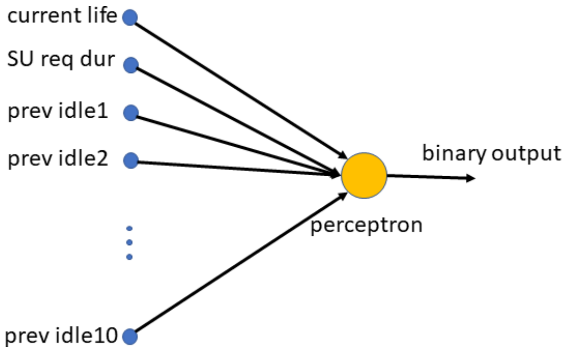


Fig. 3: Final Architecture adopted for our application. It consists of twelve node input layer connected to a single perceptron with a binay output. The perceptron uses *sigmoid* activation function, *binary cross entropy* loss function and *adam with Nesterov momentum (nadam)* optimizer.

In an earlier work, we proposed an analytical model based on *survival analysis* to design a DSA system [2]. Essentially, it computes a non-parametric estimate of survival function [24] from the idle time distribution of occupancy data. It then computes an estimate of cumulative hazard function. Given an upper bound on probablity of successful transmission, it formulates an approximate test statistic based on difference of cumulative hazard function at two different time instants (see Eqn (14) in [2]). This equation becomes the basis of two prediction algorithms (predict whether a requested SU transmission will be successful or not). We compare the performance of our ANN-based scheme, (henceforth refered to as DSA-ML scheme), with the survival analysis based scheme based on Algorithm 1 (we call DSA-SA scheme) described in [2].

G. Metrics

The following two metrics were used to evaluate the performance of DSA-ML scheme and also to compare the performance of DSA-ML scheme with DSA-SA scheme.

- White Space Utilization (WSU): White Space Utilization (WSU) of an SU on a given channel is the fraction of total idle time (or white space) used by the SU to transmit its data. It is the ratio of total idle duration used for transmission by an SU on a given channel to the total idle duration on that channel.
- Probability of Interference (PoI): The Probability of Interference (PoI) of an SU for a given channel is defined as the probability of that the SU's tranmission collides with a PU transmission. Hence, it can be approximited to be the ratio of the number of collisions with the PU

ANN Architecture	Training Accuracy (%)	Training Loss	Validation Accuracy (%)	Validation Loss	Test Accuracy (%)
three hidden layers, first with 20 nodes, second with 10 nodes, and third with 5 nodes	78.47	0.441	77.57	0.455	91.8
two hidden layers, first with 10 nodes and second with 5 nodes	77.41	0.456	77.14	0.461	95.9
one hidden layer with 8 nodes	76.95	0.463	75.93	0.469	95.9
one hidden layer with 4 nodes	76.82	0.468	75.46	0.481	95.9
one hidden layer with 2 nodes	76.36	0.479	76.51	0.478	95.9
no hidden layer (single perceptron)	74.96	0.500	74.99	0.500	95.9

TABLE I: Experiment results with different ANN architectures. Hidden layers were fully connected and used *RELU* activation function, whereas the output node used *sigmoid* activation function. *binary cross entropy* loss function and *adam with Nesterov momentum (nadam)* optimizer were employed in the ANNs. Training was stopped after 100 epochs.

	NIST Dataset		UPENN Dataset	
	day1	day2	hh1	hh2
mean idle dur (ms)	80.11	45.31	10.78	13.37
stdev idle dur (ms)	373.58	97.60	10.78	13.39
% of time channel idle	98.69	97.75	90.51	92.41

TABLE II: Some statistics of idle duration in the Datasets

transmissions to the total number of SU transmission over a very long observation period.

V. Results

In this section we present results of our experiments. We have two sets of results, one using NIST dataset and the other using UPENN dataset. In all our experiments, the SU requests arrive with a Poisson distribution and the mean interarrival time of the request is varied. Every SU request is for transmitting for a constant duration of 2 ms (i.e., τ = 2 ms). For NIST dataset we used the data corresponding to channel 5, whereas for UPENN data we use the data of channel 21. Some statistics of idle durations for the two datasets are provided in Table II.

A. Performance Evaluation using NIST Dataset

Figure 4 compares WSU of DSA-ML and DSA-SA schemes as mean interarrival time of SU request increases. In this case, the occupancy data of first day (day1) was used for training (for both DSA-ML and DSA-SA schemes) whereas the data of second day (day2) was used for testing in DSA-ML scheme and for running Algorithm1 in DSA-SA scheme, i.e., this uses configuration NIST_day1_day2. For both the schemes, WSU decreases as the mean SU request inter-arrival time increases. As mean SU request inter-arrival time increases, the number of SU request in the observation period decreases. Since the duration of SU transmission is constant (2 ms), the amount of white space utilized for SU transmission decreases. We also notice that WSU for DSA-ML scheme is always higher than that of DSA-SA scheme.

Figure 5 depicts the same performance comparision as Figure 4, except that day2 data was used for training whereas day1 data was used for testing (configuration NIST_day2_day1). In this case also we see very similar results. WSU for DSA-ML scheme is always higher than that of DSA-SA scheme. So, DSA-ML scheme outperforms DSA-SA scheme in both the configurations. DSA-SA scheme assumes that the channel idle times are independent and uses a non-parametric estimate of cumulative hazard function in its algorithms. This assumption of independence of idle times and approximation of cumulative hazard function adversely affect its performance. DSA-ML scheme is able to overcome these limitations to some extent by training its neuron on the large training dataset. Hence, DSA-ML scheme performs better than DSA-SA scheme. WSU for both the schemes do not change much between NIST_day1_day2 and NIST_day2_day1 configurations. This signifies that both the DSA schemes are robust against changing training dataset.

Figures 6 and 7 show the variation of PoI as mean SU request inter-arrival time increases when configurations NIST_day1_day2 and NIST_day2_day1 are used respectively. In configuration NIST_day1_day2, PoI for DSA-ML scheme is slightly lower than that of DSA-SA scheme and for both the schemes the PoI does not vary much as mean SU request inter-arrival time increases. This is a useful property which implies that the interference to the PUs will not vary much when the SU request arrival rate changes. The PoI for DSA-ML scheme is 0.04 which is pretty low. With NIST_day2_day1, PoI is even lower for both the schemes and does not vary much with increasing mean SU request inter-arrival time. Although PoI for DSA-ML scheme is slightly higher than that of DSA-SA scheme, it is only 0.023, which is very low.

It is worth noting that at mean SU request inter-arrival time of 2 ms the WSU for DSA-ML scheme for either configuration is above 45 % while incurring very low PoI (0.041 for NIST_day1_day2 configuration and 0.023 for NIST_day2_day1 configuration). This indicates that our DSA-

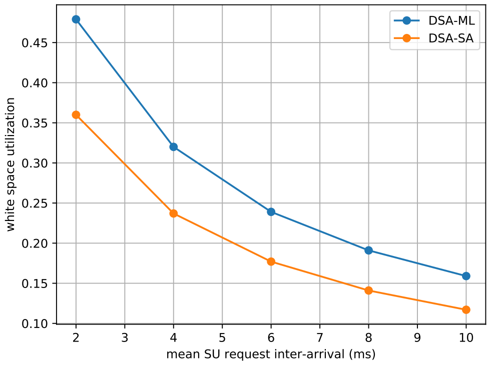


Fig. 4: WSU vs inter-arrival time (config NIST_day1_day2)

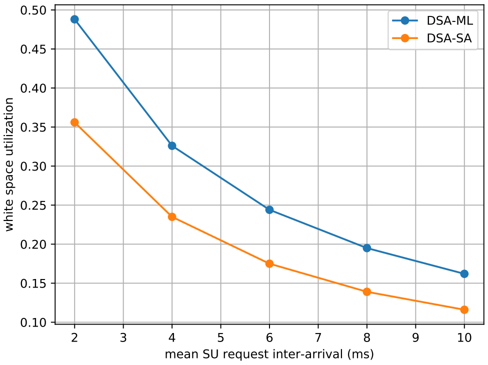


Fig. 5: WSU vs inter-arrival time (config NIST_day2_day1)

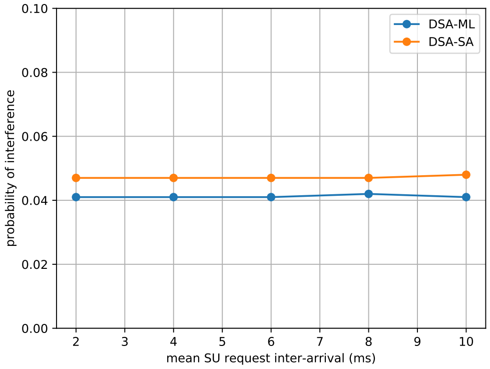


Fig. 6: PoI vs inter-arrival time (config NIST_day1_day2)

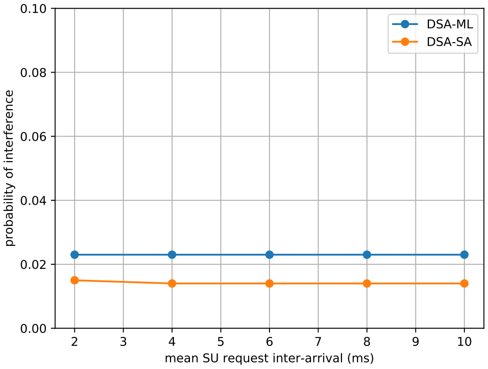


Fig. 7: PoI vs inter-arrival time (config NIST_day2_day1)

ML scheme can be used to achieve reasonably high WSU with a low PoI.

B. *Performance Evaluation using UPENN Dataset*

There is no performace numbers available for DSA-SA scheme with UPENN dataset. Hence, for this dataset we only present the results with DSA-ML scheme. Figure 8 shows the variation of WSU as mean SU request inter-arrival time increases for upenn_hh1_hh2 and upenn_hh2_hh1 configuration. As expected, WSU decreases as mean SU request inter-arrival time increases. We notice very little difference between WSU values across the two configurations, which again shows the robustness of the DSA-ML scheme for this dataset. Figure 9 shows the variation of PoI as mean SU request inter-arrival time increases for upenn_hh1_hh2 and upenn_hh2_hh1 configurations. For both the configurations PoI remains almost constant.

VI. Conclusions and Future Work

We devised a DSA scheme based on feed forward ANN which can be used in a system in which spectrum occupancy of PUs can be modelled as a sequence of busy and idle periods. We experimented with different feed forward ANN architectures for opportunistic DSA using real world LTE uplink data. Our experiments showed that a simple perceptron based ANN produces good performance for the DSA scheme. The dataset has high spectrum availability, i.e., percentage of idle duration was high and the SU request was for short duration (2 ms). So, we believe that a simple ANN architecture is adequate when spectrum has high availability and SU requests are relatively short. We compared our ML based scheme (DSA-ML) with a traditional analytical model based scheme (DSA-SA) that uses survival analysis. Our DSA-ML model performed better than DSA-SA model in terms of WSU in all the configurations. In terms of PoI, DSA-ML model sometimes performed better and some other times it was slightly worse than the DSA-SA model.

In terms of future work, we want to experiment with different ANN architectures when the spectrum availability may not be as high as in our collected dataset. For that, we may have to generate synthetic data. We also want to try a support vector machine (SVM) based model with the collected dataset

Sahoo, Anirudha. "A Machine Learning Based Scheme for Dynamic Spectrum Access." Presented at 2021 IEEE Wireless Communications and Networking Conference (WCNC 2021), Nanjing, CN. March 29, 2021 - April 01, 2021.

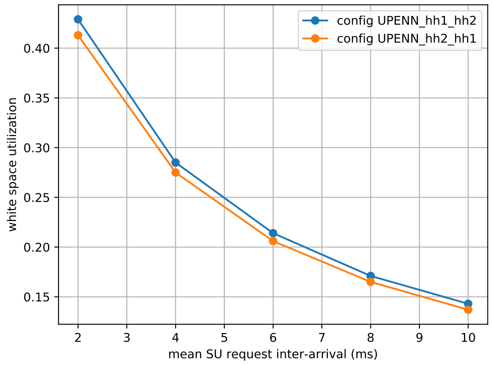


Fig. 8: WSU vs inter-arrival time using UPENN dataset in DSA-ML Scheme

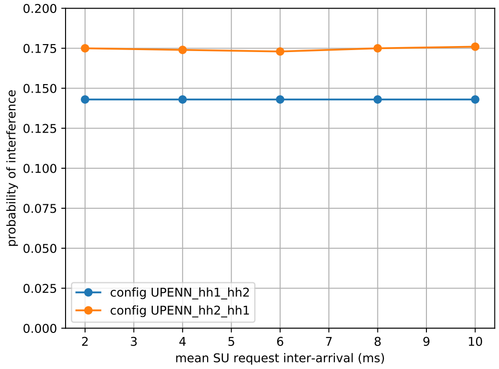


Fig. 9: PoI vs inter-arrival time using UPENN dataset in DSA-ML Scheme

and observe its performance. In this work, an SU request was only in one dimension: time. We would like to extend it to serve SU request in two dimensions: time and frequency. This, in LTE parlance, is a request for physical resource block (PRB). The SU would request for a certain number of PRBs and we need to build an ANN which can predict whether to grant or deny the request. If the request is granted, then the ANN should provide the map of the allocated PRBs.

References

[1] K. S. Tugba Erpek and D. Jones, *Dublin Ireland Spectrum Occupancy Measurements Collected On April 16-18, 2007*, 2007 (accessed September 21, 2020), http://www.sharedspectrum.com/wp-content/uploads/Ireland_Spectrum_Occupancy_Measurements_v2.pdf.

[2] T. A. Hall, A. Sahoo, C. Hagwood, and S. Streett, "Dynamic spectrum access algorithms based on survival analysis," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 740–751, Dec 2017.

[3] Q. Zhao, L. Tong, A. Swami and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 589–600, April 2007.

[4] K. W. Sung, S. Kim and J. Zander, "Temporal Spectrum Sharing Based on Primary User Activity Prediction," *IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3848–3855, December 2010.

[5] A. Plummer, M. Taghizadeh and S. Biswas, "Measurement based bandwidth scavenging in wireless networks," *IEEE Transactions on Mobile Computing*, vol. 11, no. 1, pp. 19–32, January 2012.

[6] M. Sharma and A. Sahoo, "Stochastic Model Based Opportunistic Channel Access in Dynamic Spectrum Access networks," *IEEE Transactions on Mobile Computing*, vol. 13, no. 7, pp. 1625–1639, July 2014.

[7] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, August 2012.

[8] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, December 2011, pp. 1–6.

[9] S. Yin, D. Chen, Q.Zhang, M. Liu and S. Li, "Mining Spectrum Usage Data: A Large-Scale Spectrum Measurement Study," *IEEE Transactions on Mobile Computing*, vol. 11, no. 6, pp. 1033–1046, June 2012.

[10] P. Huang, C-J. Liu, X. Yang, L. Xiao and J. Chen, "Wireless Spectrum Occupancy Prediction Based on Partial Periodic Pattern Matching," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 7, pp. 1925–1934, July 2014.

[11] A. Sahoo, T. A. Hall, and C. Hagwood, "Optimal dynamic spectrum access scheme for utilizing white space in lte systems," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–8.

[12] C. Lv, J. Wang, F. Yu, and H. Dai, "A q-learning-based dynamic spectrum allocation algorithm," in *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering*. Atlantis Press, 2013.

[13] L. R. Faganello, R. Kunst, C. B. Both, L. Z. Granville, and J. Rochol, "Improving reinforcement learning algorithms for dynamic spectrum allocation in cognitive sensor networks," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013, pp. 35–40.

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[15] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2087–2091.

[16] H. Rutagemwa, A. Ghasemi, and S. Liu, "Dynamic spectrum assignment for land mobile radio with deep recurrent neural networks," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[17] H. Jiang, H. He, and L. Liu, "Dynamic spectrum access for femtocell networks: A graph neural network based learning approach," in *2020 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2020, pp. 927–931.

[18] Y.-F. Li, H.-W. Zha, and Z.-H. Zhou, "Learning safe prediction for semi-supervised regression," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[19] M. D. Dyer, T. Murali, and B. W. Sobral, "Supervised learning and prediction of physical interactions between human and hiv proteins," *Infection, Genetics and Evolution*, vol. 11, no. 5, pp. 917–923, 2011.

[20] R. Caromi and M. Souryal, "Detection of incumbent radar in the 3.5 ghz cbrs band using support vector machines," in *2019 Sensor Signal Processing for Defence Conference (SSPD)*. IEEE, 2019, pp. 1–5.

[21] M. Lopez-Benitez and F. Casadevall, "Methodological aspects of spectrum occupancy evaluation in the context of cognitive radio," in *2009 European Wireless Conference*, May 2009, pp. 199–204.

[22] S. Roy, K. Shin, A. Ashok, M. McHenry, G. Vigil, S. Kannam, and D. Aragon, "Cityscape: A metro-area spectrum observatory," in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, July 2017, pp. 1–9.

[23] *Why is it common in Neural Network to have a decreasing number of neurons as the Network becomes deeper*, 2017 (accessed August 15, 2020), https://www.quora.com/Why-is-it-common-in-Neural-Network-to-have-a-decreasing-number-of-neurons-as-the-Network-becomes-deeper.

[24] R. G. Miller Jr, *Survival analysis*. John Wiley & Sons, 2011, vol. 66.

Characterization of residential circuit impedance

M.A. Ehsan[1,2], W. Guo[1], D.M. Anand[1] and A.M. Gopstein[1]
[1]National Institute of Standards and Technology, Gaithersburg, MD, USA
[2]University of the District of Columbia, Washington, DC, USA
{**mdamimul.ehsan**, wenqi.guo, dhananjay.anand, avi.gopstein}@nist.gov

***Abstract*—Electrical harmonics associated with switching power electronics have been observed to induce electromagnetic interference with equipment sharing the same electrical circuit. Some research studies have documented malfunction cases and even damage to particularly sensitive equipment due to this interference. The harmonic spectra of commonly used switching power electronics in a residential electrical system have been profiled before; this paper provides a phenomenological characterization of the voltage and current transfer function (in the frequency range 20 Hz – 100 kHz) from purported sources of electrical harmonics to the point at which a utility meter is typically installed (also the PCC to the service transformer). This work also evaluates the impact of shunt capacitance close to the point of coupling, resulting in a complex zero in the transfer function and the consequent impact on phase margins of the system.**

***Index Terms*—impedance measurement, transients, NZERTF, SPICE model, shunt capacitance**

I. Introduction

The advent of energy-efficient appliances and energy optimization technologies in residential and commercial buildings has created the opportunity for intelligent control of the net load while also providing ancillary functions such as voltage support and frequency ride through to the distribution network [1]. While commercial customers have historically supported the bulk of ancillary functions in the form of power quality remediation and voltage support [2], residential customers have played a somewhat passive role [3]. There is a significant interest to include residential customers in future ancillary market [4] in addition to their already expanding role in distributed generation [5]. Enabling residential consumers to actively participate in power quality improvement at the grid edge is a key step towards a decentralized regulation architecture for the electrical power system [6].

Residential scale Distributed Energy Resources (DERs) largely rely on switching electronic power converters to mediate their interaction with the electric grid. This paper intends to address some of the engineering and reliability concerns related to residential scale DERs. Some of these concerns are enumerated below:

1) Prior studies have outlined concerns that switching power converters may introduce ‘superharmonic’ distortion into the distribution network when used for regulation [7], [8]. Modulation algorithms such as nonlinear pulse shaping [9] used in current control of solid state lighting and space vector switching control [10] used in variable frequency drives have been identified as potential sources of these distortions.
2) There is a limited understanding about the additive deleterious effects of multiple switching power converters or loads connected ‘behind the meter.’ Individual devices may meet presently enforced limits on electromagnetic interference, but unforeseen dynamic properties of the circuit coupling devices to each other may amplify higher order harmonics [11].
3) A more immediate question is the impact harmonic distortions (observed over the frequency range of 2 kHz – 150 kHz) may have on electronic energy meters. Meter accuracy is susceptible to harmonic distortion to a varying degree depending on the standards applicable to the meters being tested [12], [13], although accuracy can be assured when the distortions are understood [14].

Tests conducted at the National Institute of Standards and Technology (NIST) on a range of residential energy meters meeting the ANSI C12.20 standard [15] against a circuit comprised of electronic dimmer switches and semiconductor based light sources showed minimal measurement errors $\leq 4\%$ [16]. The recorded errors, while small, do indicate algorithmic limitations to characterize in- and out-of-phase components of power in presence of harmonic distortion. While electronic energy meters are currently used for only revenue computation, it is possible that future decentralized control architectures of power system would require meters at the service entrance to provide measurements for the active control and regulation of DERs– requiring improved accuracy even under dynamic conditions [17].

This research aims to find a transfer function to characterize the propagation of voltage and current harmonics from their points of origin to measurement points or points of interconnection where controls on interference are typically enforced. It fills a gap between the measurement of harmonics at a device level and the modeling of distortion propagation at the level of a distribution circuit. A validated transfer function founded upon accurate measurements is essential to

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. 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.

Parts of this paper may have been presented in technical seminars and included in government publications. Recorded versions of those seminars and copyright free versions of publications are available at https://www.nist.gov/publications.

Anand, Dhananjay; Ehsan, Md Amimul; Guo, Wenqi. "Characterization of residential circuit impedance." Presented at Twelfth Conference on Innovative Smart Grid Technologies North America (IEEE ISGT NA 2021). February 16, 2021 - February 18, 2021.

designing control algorithms. Ideally, this transfer function would be validated across the entire frequency domain of interest, support the composition of circuit elements, and retain physical interoperability where possible. This paper describes the initial work in developing such a method.

Background material on sources of distortion, metrology challenges, and prior modeling efforts is presented in Section II of this paper. Section III presents instrumentation calibration and fixture compensation procedures. Measurements on a few residential circuits are described in Section IV. Section V outlines the lumped parameter model recovered from the collected data as well as some preliminary analyses to demonstrate how this model might be used to address the concerns discussed above.

II. Background

This section briefly reviews the work being done by others to characterize the variety of nonlinear loads and switching devices contributing to harmonic distortion in a residential circuit (as in [18]), and in modeling the propagation of distortions from within the residential circuit into distribution networks (as in [19]).

A. Distortions observed from residential loads

Measuring device-specific harmonic distortion introduced by residential loads appears to be an active area of interest. Prior work ranges from common household appliances such as lights, televisions, computers, refrigerators, laundry machines and air conditioners [7], [8] to more recent appliances such as electric vehicle chargers, residential photovoltaic systems and variable frequency drives [20]. There has been significant interest in assessing the collective harmonic characteristics and grid impact of several connected appliances [21]. There are some existing standards related to device level harmonic distortion and electromagnetic interoperability. Table I lists some studies that compared results against such standards. However, most standards, with the exception of IEC 61000-4-19 [22], seem to omit recent concerns in the frequency range 2 kHz – 150 kHz.

B. Prior circuit modeling efforts

At the distribution circuit level, grid edge loads with switching power converters are often studied as sources of abnormal behavior [26] or as systems compromising power quality [7]. Modeling efforts reflect these concerns and focus on the deleterious effects of under-damped circuit oscillation at superharmonic frequencies [27]. These models generally adopt lumped parameter estimates for interconnections due to the analytical convenience in studying transient behavior [28] and in composing such circuits into networks of oscillators [29]. Resistance, Capacitance and Inductance (R, L, and C) are the most commonly used lumped parameters so as to leverage existing circuit simulation tools [30]. In addition, some residential networks have been shown to be adequately modeled as series RL circuits [31].

TABLE I
Harmonic Standards in Literature

Reference	Frequency Range (Hz)	Measurement Location	Referenced Standard
[23]	60 – 900	PCC[a]	
[24]	60 – 660	RL[b]	IEEE-519-1992
[21]	60 – 900	RL	IEC 61000-3-2: 2005
[19]	60 – 660	RL	IEEE-519-1992
[20]	60 – 780	RL	
[7]	60 – 1500	PCC	EN 50160:2010
[8]	50 – 2500	RL	IEEE-519-2014
[25]	60 – 900	RL	IEC 61000-3-2:2014
[13]	3 k – 150 k	Inverter	IEC 61000-4-19:2014

[a] PCC = Point of common coupling
[b] RL = Residential loads

III. Experimental Setup

The first step to determine a parametric model (discussed in Section II-B) is the measurement of the complex two-port impedance of several archetypal residential circuits over the frequency range of 20 Hz – 100 kHz (next steps in the future will cover higher order frequencies). Residential circuit wiring is typically comprised of branches of 10 – 14 gauge copper conductors. The resistivity of these cables is very low (on the order of a 10^{-6} Ω/m), requiring careful design of the experimental setup to ensure compensation for the test fixture and calibration of the instrument. Moreover, calibration and compensation must be validated over 20 Hz – 100 kHz to ensure that the relative phase deviations induced by reactive elements were accurately recorded. Fig. 1 provides a schematic overview of the proposed experimental setup showing LCR meter, test fixture and circuit under test. The test fixture is annotated with 'phantom' components signifying the residual impedance and stray admittance within the fixture itself.

A. Instrumentation details

The measurement system comprises of a four-terminal precision LCR meter [32] that can measure impedance using signal levels of 50 μA – 20 mA. These signals are applied as reference sinusoidal excitation from the high side current terminal H_c. Vector measurements (magnitude and phase angle) of current and voltage drop are made at L_c, and between H_p and L_p, respectively. Four independent current-carrying and voltage-sensing leads are used between point of measurement and LCR meter terminals to minimize the lead and contact resistance effects. The impedance of the current detection path between H_c and L_c can be independently controlled using a balanced buffer between L_c and the signal ground. Vector measurements at 377 discrete frequency points (20 Hz – 100 kHz) were collected by multiple test runs and logged for analysis.

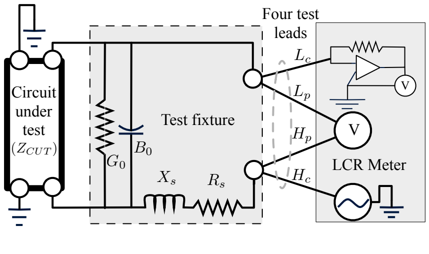


Fig. 1. Schematic diagram showing the LCR meter and the two port circuit under test (Z_{CUT}). The diagram also shows the stray admittance and residual impedance in the test fixture used in the experiment.

Anand, Dhananjay; Ehsan, Md Amimul; Guo, Wenqi. "Characterization of residential circuit impedance." Presented at Twelfth Conference on Innovative Smart Grid Technologies North America (IEEE ISGT NA 2021). February 16, 2021 - February 18, 2021.

B. Fixture compensation

The LCR meter described above was connected to the circuit under test (CUT) using four BNC terminated 5 m long RG-142 coaxial cables. These 50 Ω cables are doubly shielded and rated up to 4 GHz. The lengths were sufficient to span the distance between the point of common coupling (PCC) and test points. The cables and the interconnection hardware required to connect them to the CUT together comprised the test fixture for the experiment.

The use of four independent test leads minimizes the impact of the distributed capacitance of the test fixture. One of the two shield conductors were wired as a guard to further reduce parasitic capacitive losses at higher frequencies. In addition to passive mitigation, a compensation model for stray admittance (Y_0) and residual impedance (Z_s) of the test fixture is developed at each of the 377 measurement points. Note that Y_0 is comprised of the shunt terms: conductance (G_0) and susceptance (B_0), while Z_s is comprised of series resistance (R_s) and reactance (X_s). Fig. 1 illustrates our equivalent circuit model for the test fixture.

Since $Z_s \ll \frac{1}{Y_0}$, Z_s was measured by shorting the fixture at the point of connection to the circuit, similarly, Y_0 was measured by leaving the connection points of test fixture open. These measurements were repeated three times for all 377 test frequencies. Every measurement was then corrected to remove fixture effects using the algebraic relation (1). Where, Z_{CUT} is corrected impedance and Z_m is measurement reported by the instrument.

$$Z_{CUT} = \frac{Z_m - Z_s}{1 - Y_0(Z_m - Z_s)} \tag{1}$$

To verify the compensation model, 10 Ω, 1 nF and 10 mH reference devices are measured using the experimental setup shown in Fig. 1, and applied the aforementioned fixture compensation. Measurements of the three reference standards over the frequency range 20 Hz to 10 kHz showed a standard deviation of 1 mΩ, 3 pF and 0.5 μH respectively (corresponding to a 99 % measurement confidence of 10 Ω ± 3 mΩ, 1 nF ± 8 pF and 10 mH ± 1 μH). Note that while the capacitor and inductor used for these measurements were laboratory grade reference standards with manufacturer rated parameter variation in the order of $\pm 10^{-6}$, the resistor was a commercial-grade precision resistor rated for $\pm 10^{-3}$ variation. This verification exercise confirms that the combination of careful design of test setup and data correction for frequency-dependent changes resulted in measurements within the order of uncertainty of the reference device over a range of test frequencies. This verification step ensured that the measurement fixture used for this experiment did not contribute to a significant increase in measurement noise.

IV. Impedance Measurements of Circuit Elements

The aforementioned test fixture, instrumentation and compensation scheme were deployed to measure the impedance of three residential circuit archetypes in the Net-Zero Energy Residential Test Facility (NZERTF) located at the NIST campus in Gaithersburg, MD [33]. Three test circuits (henceforth referred to as C8, C14 and C15) were selected, all representing segments of the residential electrical network between the load connection points inside the house and PCC to the utility managed distribution network. The point of connection to the utility energy meter is PCC (Fig. 2). The load connection point was identified as plane of reference for the vector measurements taken at PCCs. There were no loads connected to the circuit during measurements. Upstream of the utility meter is a 75 kVA, 480 V to 240 V transformer providing split phase (± 120 V) service.

C14 and C15 correspond to circuit segments connecting two 120 V receptacles in a single room of the NZERTF to the PCC, while C8 corresponds to a receptacle in a utility closet. C14 and C15 are connected to two different phases of the split-phase circuit, while C8 and C15 share the same phase.

The magnitude and phase of the measured complex impedance for C8, C14 and C15 are shown in Fig. 3. These measurements are obtained after first isolating the circuits being measured from all of the other wiring in the NZERTF, the PCC was also disconnected from the secondary of the service transformer. A qualitative assessment of the frequency-dependent impedance shows that C14 and C15 have similar frequency response curves which is intuitive given their similarity in length and topology despite being connected to different phases. The series resistance of C8 is close to double that of C14 and C15 (as observed at low frequencies), but the difference in impedance magnitude between all the circuits tends to diminish as the circuit's inductive components dominate at higher frequencies. The secondary winding of the service transformer was isolated from the rest of the residential circuit to measure coil impedance. These measurements are overlaid on our circuit impedance measurements in Fig. 3 to highlight the dominantly inductive loading inherent to transformer windings.

All measured circuits eventually exhibit a 10 Ω/decade increase in magnitude and a phase lag of $\pi/2$ radians, consistent with the frequency-dependent impedance of a series RL circuit. This qualitative assertion is corroborated by literature cited in Section II.

V. Lumped Equivalent Model and Analysis

This section describes the model fitting for a series RL circuit, with corner frequency 1 kHz, to the measurements we obtained. Accordingly, the optimal estimate for the parameter

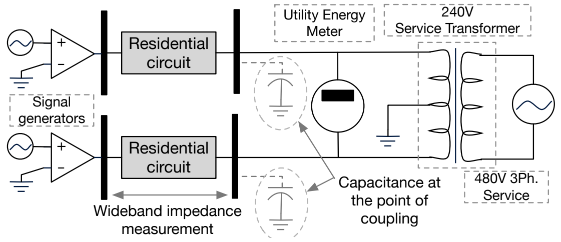


Fig. 2. A schematic view of the measurement setup used to obtain our field test data. The diagram shows the two points between which we measured impedance. The diagram also shows points at which we assume harmonic frequencies are injected into the circuit, the connection point for a service transformer, and the shunt capacitance considered in our analysis.

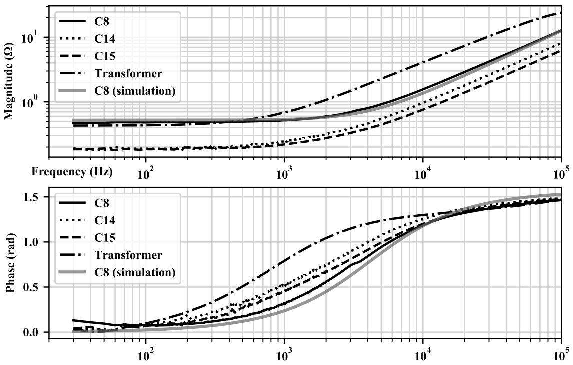


Fig. 3. Impedance measurements of C8, C14, C15 and the secondary winding of the service transformer. Also shown, simulated frequency response of a series RL circuit with parameters fit to C8.

vector $[\hat{R},\hat{L}]^T$ minimizes the two-domain least squares formulation in (2).

$$\mathbf{J}\left(\begin{bmatrix}\hat{R}\\ \hat{L}\end{bmatrix}\right)=\left\|\begin{bmatrix}|Z|_{1:m}\\ |Z|_{m+1:n}\end{bmatrix}-[\Theta]\begin{bmatrix}\hat{R}\\ \hat{L}\end{bmatrix}\right\|^2 \qquad (2)$$

The two domains correspond to m measurements of $|Z(f)|$ where $f \leq 1$ kHz and the remaining $n-m$ measurements of $|Z(f)|$ where $f > 1$ kHz, respectively. Both domains can be represented by a composite data matrix Θ of the form:

$$\Theta_{n\times 2}=\begin{bmatrix}1 & \cdots & 1_m & 0_{m+1} & \cdots & 0_n\\ 0 & \cdots & 0_m & f_{m+1} & \cdots & f_n\end{bmatrix}^T\begin{bmatrix}\frac{2\pi}{n-m} & 0\\ 0 & \frac{1}{m}\end{bmatrix}$$

Parameter estimates $[\hat{R},\hat{L}]^T$ and the corresponding mean squared error ($\mathbf{J}/n$) for each of the circuits tested are presented in Table II. Fig. 3 overlays a simulated frequency response for the impedance model of circuit C8 over the measurements obtained from C8 in order to illustrate behavioral parity between model and data in magnitude and phase regimes over the entire frequency range of interest.

A. Preliminary analyses using the equivalent model

The lumped equivalent model for circuits provides several analytical benefits. One clear benefit is– the model composition of different circuit elements using Kirchoff's Laws as constraints. This capability allows to consider and analyze hypothetical circuit configurations such as the presence of a shunt capacitance at the PCC, as illustrated in Fig. 4.

In a practical sense, this capacitance may be associated with a large reactive load or the output filter capacitor bank used with most switching inverters. A shunt capacitor placed between the inductive elements of a residential circuit and the secondary winding of the service transformer would constitute a tank circuit which might either self resonate or be excited into a sustained forced oscillation. In either case, it would be useful to determine the transfer function of the resulting circuit and perform parametric studies to evaluate the gain of transfer function from the point at which harmonics are introduced into the circuit to the PCC.

TABLE II
LUMPED PARAMETER ESTIMATES ASSUMING A SERIES RL CIRCUIT

CUT	$\hat{\mathbf{R}}$ (mΩ)	$\hat{\mathbf{L}}$ (μH)	$\mathbf{J}/n$
C8	496	20.3	0.14
C14	212	12.9	0.08
C15	200	9.9	0.07
Transformer	532	37.9	1.9

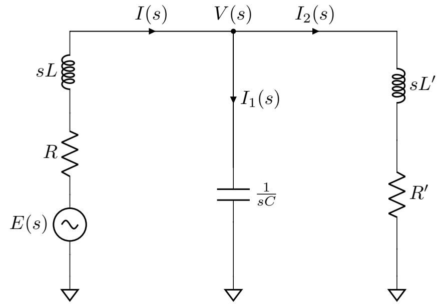


Fig. 4. An equivalent circuit for a shunt capacitance located between a residential network and the secondary winding of the service transformer.

In the equivalent circuit in Fig. 4, [R', L'] and [R, L] are parameters for the models of one leg of the center tapped transformer winding and circuit C8 respectively (Parameters noted in Table II).

When a shunt capacitance C is added to the circuit, the harmonic source E experiences a frequency dependent coupling impedance Z_{eq}, analytically derived below:

$$Z_{eq}(s)=\frac{s^3(LL'C)+s^2(RL'C+R'LC)+s(RR'C+L+L')+(R+R')}{s^2(L'C)+s(R'C)+1}$$

Then the transfer functions for voltage V and current I_2 at the PCC in response to a harmonic voltage excitation E can be derived as follows:

$$\frac{V(s)}{E(s)}=\frac{R'+sL'}{s^3(LL'C)+s^2(RL'C+R'LC)+s(RR'C+L+L')+(R+R')}$$
$$\frac{I_2(s)}{E(s)}=\frac{1}{s^3(LL'C)+s^2(RL'C+R'LC)+s(RR'C+L+L')+(R+R')}$$

Similarly, the voltage and current responses at the PCC to a harmonic current injection I are:

$$\frac{V(s)}{I(s)}=\frac{R'+sL'}{s^2(L'C)+s(R'C)+1}$$
$$\frac{I_2(s)}{I(s)}=\frac{1}{s^2(L'C)+s(R'C)+1}$$

These transfer functions can be used to simulate the propagation dynamics of harmonics, transients, switching artifacts, etc. which have a spectral footprint between 20 Hz and 100 kHz. One such simulation study revealed that when the shunt capacitance is approximately 17.4 μF, voltage gain of V/E peaks at 14 dB @ 10 kHz.

These transfer functions can also provide insight on the stability of the resonant circuit over a range of shunt capacitance values. Table III shows the phase margin of the circuit in Fig. 4 as the value of C is increased. The phase margin of a dynamic system represents the domain of stability available when the system is enclosed in a feedback network. In the example circuit, this may be a feedback controller used by an inverter to regulate power quality. Analysis illustrates some interesting behavior: while the responses of V/I, V/E and I_2/I show phase margin improvements with increasing

TABLE III
CHANGES TO THE SYSTEM PHASE MARGIN IN RESPONSE TO INCREASING SHUNT CAPACITANCE.

Transfer Function	**Phase Margin** (dB)									
	$C = 0$	$C = 1$ pF	$C = 1$ nF	$C = 1$ μF	$C = 5$ μF	$C = 10$ μF	$C = 17.4$ μF	$C = 50$ μF	$C = 1$ mF	$C = 1$ F
V/I	-105	90	90	90	90.7	92.2	95.1	113	Inf	Inf
V/E	Inf	0.00941	0.298	9.43	21.3	30.4	40.8	78	Inf	Inf
I_2/I	-180	0.00495	0.157	4.95	11.1	15.7	20.8	35.6	-180	-180
I_2/E	140	41.6	-41.4	-34.9	-27.1	-21.5	-15.6	0.127	99.8	140

capacitance consistent with intuition, the transfer function I_2/E suggests that for range of values of C between 1 nF and 50 μF, closed loop feedback would cause unstable circuit response. An unstable circuit would effectively amplify the spurious harmonics introduced by E and potentially propagate these artifacts upstream.

VI. CONCLUSION AND FUTURE WORK

This paper presented an experimental setup complete with calibration and compensation schemes to accurately measure the impedance of residential scale electrical circuits. Measurements of selected circuit segments at the NZERTF were taken between 20 Hz and 100 KHz and lumped circuit models were developed from the data. The frequency range selected for this work was based on a survey conducted on existing standards to identify gaps.

Preliminary results are reported from the analyses conducted on derived models, studying the gain and stability margins of one test circuit under reactive loading.

Moving forward, the aim is to use the methods described here to collect measurements from several more circuit segments eventually scaling up to a model of a full residential circuit including all possible switch states and expand to higher order frequencies. A library of such models representing the wide-band response of residential circuits will be critical to ensure reliable ancillary service to the grid.

REFERENCES

[1] B. Singh, A. Chandra, and K. Al-Haddad, *Power quality: problems and mitigation techniques*. John Wiley & Sons, 2014.

[2] T. M. Blooming and D. J. Carnovale, "Application of IEEE STD 519-1992 harmonic limits," in *Proc. IEEE PPFIC*, 2006, pp. 1–9.

[3] V. L. Martin, F. J. Azcondo, and A. Pigazo, "Power quality enhancement in residential smart grids through power factor correction stages," *IEEE TIE*, vol. 65, no. 11, pp. 8553–8564, 2018.

[4] D. Holmberg and F. Omar, "Characterization of residential distributed energy resource potential to provide ancillary services," *NIST SP 1900-601*, 2018.

[5] I. I. Avramidis, V. A. Evangelopoulos, P. S. Georgilakis, and N. D. Hatziargyriou, "Demand side flexibility schemes for facilitating the high penetration of residential distributed energy resources," *IET GTD*, vol. 12, no. 18, pp. 4079–4088, 2018.

[6] A. Gopstein, C. Nguyen, C. O'Fallon, N. Hastings, and D. Wollman, "NIST framework and roadmap for smart grid interoperability standards, release 4.0," *NIST SP XXX*, 2020.

[7] J. Niitsoo, I. Palu, J. Kilter, P. Taklaja, and T. Vaimann, "Residential load harmonics in distribution grid," in *Proc. IEEE EPECS*, 2013, pp. 1–6.

[8] K. Nikum, R. Saxena, and A. Wagh, "Harmonic analysis of residential load based on power quality," in *Proc. IEEE PIICON*, 2016, pp. 1–6.

[9] H. Safamehr, T. A. Najafabadi, and F. R. Salmasi, "Enhanced control of grid-connected inverters with non-linear inductor in LCL filter," *IET PEL*, vol. 9, no. 10, pp. 2111–2120, 2016.

[10] Y. Deng and R. G. Harley, "Space-vector versus nearest-level pulse width modulation for multilevel converters," *IEEE TPEL*, vol. 30, no. 6, pp. 2962–2974, 2015.

[11] T. Rehman, J. Yaghoobi, and F. Zare, "Harmonic issues in future grids with grid connected solar inverters: 0–9 khz," in *Proc. IEEE AUPEC*, 2018, pp. 1–6.

[12] I. Diahovchenko *et al.*, "Effect of harmonic distortion on electric energy meters of different metrological principles," *Front. Energy*, vol. 13, no. 2, pp. 377–385, 2019.

[13] F. Leferink, C. Keyer, and A. Melentjev, "Static energy meter errors caused by conducted electromagnetic interference," *IEEE EMC Mag.*, vol. 5, no. 4, pp. 49–55, 2016.

[14] R. Steiner *et al.*, "A NIST testbed for examining the accuracy of smart meters under high harmonic waveform loads," *NISTIR 8248*, 2019.

[15] *American national standard for electricity meters - 0.2 and 0.5 accuracy classes*, ANSI Std. CI2.20–2015, 2017.

[16] R. Steiner, M. Farrell, S. Edwards, J. Ford, and T. Nelson, "A NIST testbed for examining the accuracy of smart meters under high harmonic waveform loads," in *Proc. IEEE CPEM*, 2018, pp. 1–2.

[17] A. Mohsenian-Rad, V. Wong, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE TSG*, vol. 1, no. 3, pp. 320–331, 2010.

[18] H. E. Mazin, E. E. Nino, W. Xu, and J. Yong, "A study on the harmonic contributions of residential loads," *IEEE TPWRD*, vol. 26, no. 3, pp. 1592–1599, 2011.

[19] Y. Wang, R. M. O'Connell, and G. Brownfield, "Modeling and prediction of distribution system voltage distortion caused by nonlinear residential loads," *IEEE TPWRD*, vol. 16, no. 4, pp. 744–751, 2001.

[20] H. Sharma, M. Rylander, and D. Dorr, "Grid impacts due to increased penetration of newer harmonic sources," *IEEE TIA*, vol. 52, no. 1, pp. 99–104, 2016.

[21] C. Jiang, D. Salles, W. Xu, and W. Freitas, "Assessing the collective harmonic impact of modern residential loads— Part II: Applications," *IEEE TPWRD*, vol. 27, no. 4, pp. 1947–1955, 2012.

[22] *Electromagnetic compatibility*, IEC Std. 61 000-4-19:2014, 2015.

[23] A. Mansoor *et al.*, "Predicting the net harmonic currents produced by large numbers of distributed single-phase computer loads," *IEEE TPWRD*, vol. 10, no. 4, pp. 2001–2006, 1995.

[24] A. E. Emanuel, J. A. Orr, D. Cyganski, and E. M. Gulachenski, "A survey of harmonic voltages and currents at the customer's bus," *IEEE TPWRD*, vol. 8, no. 1, pp. 411–421, 1993.

[25] P. Bagheri, W. Xu, and T. Ding, "A distributed filtering scheme to mitigate harmonics in residential distribution systems," *IEEE TPWRD*, vol. 31, no. 2, pp. 648–656, 2016.

[26] E. El-Saadany, "Parameters affecting harmonic propagation and distortion levels in nonlinear distribution systems," in *Proc. IEEE PES Summer Meeting*, vol. 2, 2002, pp. 1010–1016.

[27] S. Lakrih and J. Diouri, "Wide band frequency lumped parameters equivalent model for power networks," in *Proc. IEEE ICEIT*, 2016, pp. 547–552.

[28] M. M. Forti, L. M. Millanta, and C. F. M. Carobbi, "Low-frequency transients and impedance in the power mains considering line loading," *IEEE TEMC*, vol. 38, no. 3, pp. 310–317, 1996.

[29] E. Tegling, M. Andreasson, J. W. Simpson-Porco, and H. Sandberg, "Improving performance of droop-controlled microgrids through distributed pi-control," in *Proc. IEEE ACC*, 2016, pp. 2321–2327.

[30] G. Antonini, "SPICE equivalent circuits of frequency-domain responses," *IEEE TEMC*, vol. 45, no. 3, pp. 502 – 512, 2003.

[31] R. Matias, B. Cunha, and R. Martins, "Modeling inductive coupling for wireless power transfer to integrated circuits," in *Proc. IEEE WPT*, 2013, pp. 198–201.

[32] *Impedance measurement handbook*, 6th ed., Keysight Technology, 2016.

[33] P. D. Domich, B. Pettit, A. H. Fanney, and W. M. Healy, "Research and development opportunities for the NIST net zero energy residential test facility," *NIST Tech. Note 1869*, 2015.

SoK: How (not) to Design and Implement Post-Quantum Cryptography

James Howe[1], Thomas Prest[1], and Daniel Apon[2]

[1] PQShield, Oxford, UK.
{james.howe,thomas.prest}@pqshield.com
[2] National Institute of Standards and Technology, USA.
daniel.apon@nist.gov

Abstract Post-quantum cryptography has known a Cambrian explosion in the last decade. What started as a very theoretical and mathematical area has now evolved into a sprawling research field, complete with side-channel resistant embedded implementations, large scale deployment tests and standardization efforts. This study systematizes the current state of knowledge on post-quantum cryptography. Compared to existing studies, we adopt a transversal point of view and center our study around three areas: (i) paradigms, (ii) implementation, (iii) deployment. Our point of view allows to cast almost all classical and post-quantum schemes into just a few paradigms. We highlight trends, common methodologies, and pitfalls to look for and recurrent challenges.

1 Introduction

Since Shor's discovery of polynomial-time quantum algorithms for the factoring and discrete logarithm problems, researchers have looked at ways to manage the potential advent of large-scale quantum computers, a prospect which has become much more tangible of late. The proposed solutions are cryptographic schemes based on problems assumed to be resistant to quantum computers, such as those related to lattices or hash functions. *Post-quantum cryptography* (PQC) is an umbrella term that encompasses the design, implementation, and integration of these schemes. This document is a Systematization of Knowledge (SoK) on this diverse and progressive topic.

We have made two editorial choices. First, an exhaustive SoK on PQC could span several books, so we limited our study to signatures and key-establishment schemes, as these are the backbone of the immense majority of protocols. This study will not cover more advanced functionalities such as homomorphic encryption schemes, threshold cryptography, et cetera.

Second, most surveys to-date are either (i) organized around each *family* [23] – (a) lattices, (b) codes, (c) multivariate equations, (d) isogenies, (e) hash and one-way functions – or (ii) focused on a single family [146, 83]. Our study instead adopts a transversal approach, and is organized as follows: (a) paradigms, (b) implementation, and (c) deployment. We see several advantages to this approach:

Howe, James; Prest, Thomas; Apon, Daniel. "SoK: How (not) to Design and Implement Post-Quantum Cryptography." Presented at Cryptographers' Track at the RSA Conference (CT-RSA 2021). May 17, 2021 - May 20, 2021.

- Compared to previous surveys, it provides a new point of view that abstracts away much of the mathematical complexity of each family, and instead emphasizes common paradigms, methodologies, and threat models.
- In practice, there are challenges that have been solved by one family of scheme and not another. This document's structure makes it easy to highlight *what* these problems are, and *how* they were solved. Consequently, it aims to provide specific direction for research; i.e., (i) problems to solve, and (ii) general methodologies to solve them.
- If a new family of hardness assumptions emerges – as isogeny-based cryptography recently has – we hope the guidelines in this document will provide a framework to safely design, implement, and deploy schemes based on it.

1.1 Our Findings

A first finding is that almost all post-quantum (PQ) schemes fit into one of four paradigms: Fiat-Shamir signatures, Hash-then-sign, Diffie-Hellman key-exchange, and encryption. Moreover, the same few properties (e.g., homomorphism) and folklore tricks are leveraged again and again.

Successful schemes do not hesitate to *bend* paradigms in order to preserve the security proof *and* the underlying assumption. In contrast, forcing an assumption into a paradigm may break the assumption, the security proof, or both.

Our second finding is that many PQ schemes fell short in secure, isochronous implementations which in turn lead to undeserved opinions on side-channel vulnerabilities. We also find some PQ schemes are significantly more amenable to implementations in hardware, software, their efficiencies with masking, which then translates into how performant they are in various use-cases.

Our last finding (see the full version [110]) is that all real-world efforts to deploy post-quantum cryptography will have to contend with new, unique problems. They may require a diverse combination of computational assumptions *woven together* into a single hybrid scheme. They may require special attention to *physical management* of sensitive state. And they have very unbalanced performance profiles, requiring different solutions for different application scenarios.

2 The Raw Material: Hard Problems

We first present the raw material from which cryptographic schemes are made of: hard problems. Although there exists a myriad of post-quantum hard problems, many of them share similarities that we will highlight.

2.1 Baseline: Problems that are not Post-Quantum

We first present problems that are classically hard but quantumly easy. The first family of problems relates to the discrete logarithm in finite groups; that is, the Discrete Logarithm (DLOG) problem, the Decisional Diffie-Hellman (DDH), and the Computational Diffie-Hellman (CDH) problems.

Definition 1 (DLOG/DDH/CDH). *Let $\mathbb{G}$ be a cyclic group of generator g. The discrete logarithm problem (DLOG) and the decisional/computational Diffie-Hellman problems (DDH/CDH) are defined as follows:*

– ***DLOG:*** *Given g^a for a random $a \in |\mathbb{G}|$, find a.*
– ***DDH:*** *Given g^a, g^b and g^c for random $a, b \in |\mathbb{G}|$, determine if $c = ab$.*
– ***CDH:*** *Given g^a, g^b for random $a, b \in |\mathbb{G}|$, compute g^{ab}.*

In cryptography, $\mathbb{G}$ is usually the ring $\mathbb{Z}_p$ for a large prime p, or the group of rational points of an elliptic curve. The following algebraic relations are extremely useful to build cryptosystems, for example Schnorr signatures [168] use (1) and (2) whereas the Diffie-Hellman key-exchange [72] uses (2):

$$g^a \cdot g^b = g^{a+b}, \tag{1}$$

$$(g^a)^b = \left(g^b\right)^a = g^{ab}. \tag{2}$$

The second family of problems relates to factoring.

Definition 2 (RSA and Factoring). *Let p, q be large prime integers, $N = p \cdot q$ and e be an integer.*

– ***Factoring:*** *Given N, find p and q.*
– ***RSA:*** *Efficiently invert the following function over a non-negligible fraction of its inputs:*

$$x \in \mathbb{Z}_N \mapsto x^e \bmod N. \tag{3}$$

For adequate parameters, the problems in Def. 1 and 2 are believed hard to solve by classical computers. However, Shor has shown that they are solvable in polynomial time by a quantum computer [172]. As these problems underlie virtually all current public-key cryptosystems, Shor's discovery motivated the following research for alternative, quantum-safe problems.

2.2 Problems on Lattices

The most well-known problems based on lattices are Learning With Errors (LWE) [158, 134], Short Integer Solution (SIS) [2, 130] and "NTRU" [107].

Definition 3 (SIS, LWE, and NTRU). *Let $\mathcal{R} = \mathbb{Z}_q[x]/(\phi(x))$ be a ring, and $\mathbf{A} \in \mathcal{R}^{n \times m}$ be uniformly random. The Short Integer Solution (SIS) and Learning with Errors (LWE) problems are defined as follows:*

– ***SIS:*** *Find a short nonzero $\mathbf{v} \in \mathcal{R}^m$ such that $\mathbf{Av} = 0$.*
– ***LWE:*** *Let $\mathbf{b} = \mathbf{A}^t\mathbf{s} + \mathbf{e}$, where $\mathbf{s} \in \mathcal{R}^n$ and $\mathbf{e} \in \mathcal{R}^m$ are sampled from the 'secret' distribution and 'error' distribution, respectively.*
  - ***Decision:*** *Distinguish $(\mathbf{A}, \mathbf{b})$ from uniform.*
  - ***Search:*** *Find $\mathbf{s}$.*

– ***NTRU:*** *Let $h = f/g \in \mathcal{R}$, where $f, g \in \mathcal{R}$ are 'short.' Given h, find f, g.*

SIS, LWE, and NTRU exist in many variants [158, 134, 130, 150], obtained by changing $\mathcal{R}, n, m$, or the error distributions. To give a rough idea, a common choice is to take $\mathcal{R} = \mathbb{Z}_q[x]/(x^d + 1)$, with d a power-of-two, and n, m such that nd and md are in the order of magnitude of 1000. The versatility of SIS, LWE, and NTRU is a blessing and a curse for scheme designers, as it offers freedom but also makes it easy to select insecure parameters [148].

We are not aware of closed formulae for the hardness of SIS, LWE, and NTRU. However, the most common way to attack these problems is to interpret them as lattice problems, then run lattice reduction algorithms [7, 5]. For example, the BKZ algorithm [169] with a blocksize $B \leq nd$ is estimated to solve these in time $\tilde{O}(2^{0.292 \cdot B})$ classically [18], and $\tilde{O}(2^{0.265 \cdot B})$ quantumly [127] via Grover's algorithm.

2.3 Problems on Codes

Error-correcting codes provide some of the oldest post-quantum cryptosystems. These usually rely on two problems:

- The Syndrome Decoding (SD) problem, see Def. 4.
- Hardness of distinguishing a code in a family $\mathcal{F}$ from a pseudorandom one.

We first present SD. Note that it is similar to SIS (Def. 3).

Definition 4 (SD). *Given a matrix* $\mathbf{H} \in \mathbb{F}_2^{k \times n}$ *and a syndrome* $\mathbf{s} \in \mathbb{F}_2^k$*, the Syndrom Decoding (SD) problem is to find* $\mathbf{e} \in \mathbb{F}_2^n$ *of Hamming weight* w *such that* $\mathbf{He} = \mathbf{s}$.

Since 1962, several algorithms have been presented to solve the SD problem, their complexity gradually improving from $2^{0.1207n}$ [155] to $2^{0.0885n}$ [39]. These algorithms share similarities in their designs and [177] recently showed that when $w = o(n)$, they all have the same asymptotic complexity $\approx 2^{w \log_2(n/k)}$. For many of these algorithms, quantum variants have been proposed. They achieve quantum complexities that are essentially square roots of the classical ones, by using either Grover or quantum walks.

The second problem is not as clearly defined, as it is rather a class of problems. Informally, it states that for a given family $\mathcal{C} = (C_i)_i$ of codes, a matrix $\mathbf{G}$ generating a code $C_i \in \mathcal{C}$ is hard to distinguish from a random matrix. For example, two variants of BIKE [9] assume that it is hard to distinguish from random either of these *quasi-cyclic codes* (or QC codes):

$$h_0/h_1 \tag{4}$$

$$g, g \cdot h_0 + h_1 \tag{5}$$

where $g, h_0, h_1 \in \mathbb{F}_2[x]/(x^r - 1)$, g is random and h_0, h_1 have small Hamming weight. Note that (4) and (5) are reminiscent of NTRU and (ring-)LWE, respectively (see Def. 3). Hence all the lattice problems we have defined have code counterparts, and reciprocally. Besides the QC codes of (4)-(5), another popular family of codes are Goppa codes [135, 55, 24].

2.4 Problems on Multivariate Systems

The third family of problems is based on multivariate systems. In practice, only multivariate *quadratics* (i.e., of degree 2) are used. They are the Multivariate Quadratic (MQ) and Extended Isomorphism of Polynomials (EIP) problems.

Definition 5 (MQ and EIP). *Let $\mathbb{F}$ be a finite field. Let $\mathbf{F} : \mathbb{F}^n \to \mathbb{F}^m$ of the form $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, where each $f_i : \mathbb{F}^n \to \mathbb{F}$ is a multivariate polynomial of degree at most 2 in the coefficients of $\mathbf{x}$.*

- ***MQ:*** *Given $\mathbf{y} \in \mathbb{F}^m$ and the map $\mathbf{F}$:*
  - ***Decision:*** *Is there an $\mathbf{x}$ such that $\mathbf{F}(\mathbf{x}) = \mathbf{y}$?*
  - ***Search:*** *Find $\mathbf{x}$ such that $\mathbf{F}(\mathbf{x}) = \mathbf{y}$.*
- ***EIP:*** *Let $\mathbf{S} : \mathbb{F}^n \to \mathbb{F}^n$ and $\mathbf{T} : \mathbb{F}^m \to \mathbb{F}^m$ be uniformly random affine maps. Given $\mathbf{P} = \mathbf{S} \circ \mathbf{F} \circ \mathbf{T}$ and the promise that the map $\mathbf{F}$ is in a publicly known set $\mathcal{F}$, find $\mathbf{F}$.*

Note that MQ is solvable in polynomial time for $m^2 = O(n)$ or $n^2 = O(m)$; therefore this problem is more interesting when $n = \Theta(m)$, which we assume henceforth. Also note that EIP can be parameterized by the set $\mathcal{F}$ to which the secret map $\mathbf{F}$ belongs. For example, the Unbalanced Oil and Vinegar (UOV) and Hidden Field Equation (HFEv) problems, used by Rainbow [73] and GeMSS [43] respectively, are instantiations of the EIP "framework".

Algorithms for solving MQ or EIP include F4/F5 [81], XL [56, 71] or Crossbred [121]. The best algorithms [181, 30, 121] combine algebraic techniques – e.g., solving Gröbner bases – with exhaustive search, which can be sped up using Grover's algorithm in the quantum setting, see [28] as an example. The asymptotic complexities of these algorithms are clearly exponential in n, but we did not find simple formulae to express them (either classically or quantumly), except for special cases ($q = 2$ and $n = m$) which do not accurately reflect concrete instantiations such as the signature schemes Rainbow [73] and MQDSS [165].

2.5 Problems on One-Way and Hash Functions

The most peculiar family of PQ problems relates to properties of (generic) one-way and hash functions. These problems are algebraically unstructured, which is desirable security-wise, but tends to imply more inefficient schemes.

Definition 6 (Problems on hash functions). *Let $H : X \to Y$ be a function, where $Y = 2^n$.*

- ***Preimage:*** *Given $y \in Y$, find $x \in X$ such that $H(x) = y$.*
- ***Second preimage:*** *Given $x_1 \in X$, find $x_2 \neq x_1$ such that $H(x_1) = H(x_2)$.*
- ***Collision:*** *Find $x_1 \neq x_2$ such that $H(x_1) = H(x_2)$.*

The best classical algorithm against (second) preimage is exhaustive search, hence a complexity $O(2^n)$. Grover's famous quantum algorithm [97] performs this search with a quadratic speed-up, hence a complexity $O(2^{n/2})$. Regarding collision, the best classical algorithm is the birthday attack with a complexity $O(2^{n/2})$, and (disputed) results place the complexity of the best quantum attack between $O(2^{2n/5})$ [47] and $\Theta(2^{n/3})$ [184].

2.6 Problems on Isogenies

Isogeny problems provide a higher-level twist on Def. 1. Elliptic curve cryptography posits that when given g and g^a, with g being a point on an elliptic curve E, it is hard to recover a. Similarly, isogeny-based cryptography posits that given elliptic curves E and E' over $\mathbb{F}_{p^2}$, it is hard to find a surjective group morphism (or *isogeny*, in this context) $\phi : E \to E'$.

Isogeny-based cryptography is a fast-moving field. Elliptic curves can be ordinary ($E[p] \simeq \mathbb{Z}_p$) or supersingular ($E[p] \simeq \{0\}$). Recall that the torsion subgroup $E[n]$ is the kernel of the map $P \in E \mapsto [n]P$. Most isogeny schemes work with supersingular curves, which parameters scale better. Two problems (or variations thereof) have emerged. Def. 7 provides simplified descriptions of them.

Definition 7 (Problems on isogenies). *We define the Supersingular Isogeny Diffie-Hellman (SIDH) and Commutative SIDH (CSIDH) problems as follows:*

- ***SIDH:*** *Given two elliptic curves E, E_A and the value of an isogeny $\phi : E \to E_A$ on $E[\ell^e]$, find ϕ.*
- ***CSIDH:*** *Given two elliptic curves E, E_A, find an efficiently computable isogeny $\phi \in \mathcal{C}\ell(\mathcal{O})$ s.t. $E_A = \phi \cdot E$, where $\mathcal{C}\ell(\mathcal{O})$ is the class group of $\mathcal{O} = \mathbb{Z}[\sqrt{-p}]$.*

Note that the CSIDH problem adapts DDH to the isogeny setting, and one can similarly adapt CDH (see Def. 1). Note that both problems are quantumly equivalent [89], whereas CDH and DDH are not known to be classically equivalent, except in special cases.

For SIDH, the best classical attack is via a claw-finding algorithm due to van Oorschot-Wiener [178]. Surprisingly, a recent result [120] shows that the best known quantum attack performs *worse* than [178]. The hardness of CSIDH reduces to solving a hidden shift problem, for which Kuperberg proposed quantum sub-exponential algorithms [125, 126]. The actual quantum security of CSIDH is still being debated [37, 147].

2.7 Summary of Problems

Fig. 1 summarizes the classical and quantum hardness estimates of the problems we presented. Quantum estimates are particularly prone to change, notably due to (a) the lack of clear consensus on the cost of quantum memory, (b) the prospect of future algorithmic improvements.

Figure 1: Classical and quantum hardness of some problems.

Problem	Factoring /DLOG	SIS /LWE	SD	MQ	EIP	SIDH	CSIDH	(Second) Preimg.	Coll.
Classical	$e^{\tilde{O}\left((\log p)^{1/3}\right)}$	$2^{0.292 \cdot B}$	$2^{0.0885 \cdot n}$	?	?	$O(p^{1/4})$	$O(p^{1/4})$	$O(2^n)$	$O(2^{n/2})$
Quantum	poly(N)	$2^{0.265 \cdot B}$	$2^{0.05804 \cdot n}$	?	?	$O(p^{1/4})$	$e^{\tilde{O}\left(\sqrt{\log p}\right)}$	$O(2^{n/2})$	$\Theta(2^{n/3})$

Howe, James; Prest, Thomas; Apon, Daniel. "SoK: How (not) to Design and Implement Post-Quantum Cryptography." Presented at Cryptographers' Track at the RSA Conference (CT-RSA 2021). May 17, 2021 - May 20, 2021.

3 Paradigms are Guidelines, not Panaceas

In the classical world, there are two paradigms for signing:

- Fiat-Shamir (FS) [85], proven in the random oracle model (ROM) by [153]. One example is Schnorr signatures and the (Elliptic Curve) Digital Signature Algorithm, (EC)DSA.
- Hash-then-sign. The most prominent formalization of this paradigm is the Full Domain Hash [21] (FDH), proven in the ROM by [22, 54]. Numerous instantiations exist, such as RSA-PSS (Probabilistic Signature Scheme) and Rabin signatures.

There are also two paradigms for key establishment:

- Public-key encryption (PKE), like El Gamal [78] or RSA [160].
- Diffie-Hellman (DH) key-exchange [72].

At a conceptual level, this section shows that most PQ signature or key establishment schemes can be cast under one of these four paradigms. This is summarized by Table 1, which also provides us with two open questions:

(Q1) Can we have isogeny-based Hash-then-sign schemes?
(Q2) Can we have multivariate key establishment schemes?

The prospect that we will have practical key establishment schemes based on symmetric primitives only seems unlikely, see [14]. For (Q1) and (Q2), we hope that the guidelines provided in this section will help to answer them. Our main takeaway is that scheme designers should treat paradigms as guidelines. In particular, a fruitful approach is to weaken some properties, as long as the final scheme achieves meaningful security notions. For example:

Table 1: Correspondence between post-quantum schemes and problems.

	Signature		Key Establishment	
	Hash-&-Sign	Fiat-Shamir	DH-style	PKE
Lattices	[156, 50]	[133, 36]	[149]	[170, 61, 185]
Codes	[68]	[174, 180]	[24, 9]	[1]
Isogenies	?	[65, 34]	[45]	[117]
Multivariate	[73, 43]	[165]	?	?
Symmetric	[115]	[183, 32]	-	-

- Efficient PQ variants of the FDH framework discards trapdoor permutations for weakened definitions, which suffice for signatures, see Sec. 3.3.
- *Fiat-Shamir with Aborts* changes the protocol flow and may only prove knowledge of an approximate solution. This suffices for signatures, see Sec. 3.1

On the other hand, designers should not try to cram a problem into a predefined paradigm, as it often results in impractical (if not broken) parameters. Examples are rigid adaptations of:

- DH with lattices [102] and isogenies [66], see Sec. 3.4.
- FDH with codes [55] or lattices [105], see Sec. 3.3.

3.1 Schnorr Signatures over Lattices

Fig. 2 recalls the structure of an identification scheme, or ID scheme. Any ID scheme can be converted into a signature via the Fiat-Shamir transform [85]. A efficient ID scheme is Schnorr's 3-move protocol [168]. It instantiates Fig. 2 with the parameters in Table 2 (column 2). It also requires additive and multiplicative properties similar to (1)-(2).

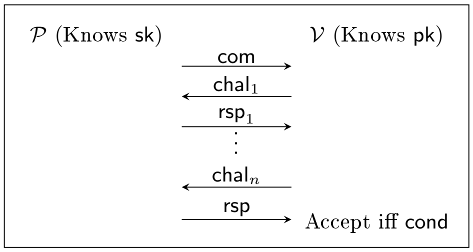

$\mathcal{P}$ (Knows sk)
$\mathcal{V}$ (Knows pk)
com
chal_1
rsp_1
$\vdots$
chal_n
rsp
Accept iff cond

Figure 2: A $(2n+1)$-move ID scheme.

$$\begin{tikzcd}
E \arrow[r, "\phi_{\mathsf{com}}"] \arrow[d, "\phi_{\mathsf{sk}}"] & E_{\mathsf{com}} \arrow[d, "\phi_{\mathsf{chal}}"] \\
E_{\mathsf{pk}} \arrow[r, "\phi_{\mathsf{rsp}}"] & E_{\mathsf{chal}}
\end{tikzcd}$$

Figure 3: SQISign

Fortunately, lattice and code problems do have properties similar to (1)-(2). An early attempt to propose Schnorr lattice signatures is NSS (NTRU-based Signature Scheme) [106], which was broken by statistical attacks [92]. The high-level explanation is that the ID scheme in NSS did not satisfy the *honest verifier zero-knowledge* (HVZK) property. Each transcript leaked a bit of information about sk, which [92] exploited to recover sk. This was fixed by Lyubashevsky's scheme [132], by giving the prover the possibility to abort the protocol with a probability chosen to factor out the dependency to sk from the signature. This changes the flow of the ID scheme, but allows to prove HVZK. It is also invisible to the verifier as the signer will simply restart the signing procedure in case of an abort. An example instantiation is shown in Table 2 (column 3).

On the other hand, properties of lattices enable specific tricks tailored to this setting. For example, for LWE, least significant bits (LSBs) do not really matter. Let $\lfloor \mathbf{u} \rceil_b$ be a lossy representation of $\mathbf{u}$ that discards the b LSBs for each coefficient of $\mathbf{u}$. Finding a search-LWE solution $(\mathbf{s}_1, \mathbf{s}_2)$ for $(\mathbf{A}, \lfloor \mathbf{t} \rceil_b)$ implies a solution $(\mathbf{s}_1, \mathbf{s}'_2)$ for $(\mathbf{A}, \mathbf{t})$, with $\|\mathbf{s}_2 - \mathbf{s}'_2\|_\infty \leq 2^b$. This indicates that, as long as b is not too large, LSBs are not too important for LWE.

This intuition was formalized by [13], who show that dropping $\mathbf{z}_2$ and checking only the high bits of com allowed to reduce the signature size by about 2, for essentially the same (provable) security guarantees. Similarly, [98] applied this idea to reduce the public key size. The idea was improved upon by

Table 2: Instantiations of Schnorr Signatures.

Element	Schnorr	Lyubashevsky (w/ LWE)
sk	Uniform x	Short $(\mathbf{s}_1, \mathbf{s}_2)$
pk	$g, h = g^x$	$\mathbf{A}, \mathbf{t} = \mathbf{A} \cdot \mathbf{s}_1 + \mathbf{s}_2$
com	g^r for uniform r	$\mathbf{A} \cdot \mathbf{r}_1 + \mathbf{r}_2$ for short $(\mathbf{r}_1, \mathbf{r}_2)$
chal	Uniform c	Short c
rsp	$r - cx$	$(\mathbf{z}_1, \mathbf{z}_2) = (\mathbf{r}_1 - c\mathbf{s}_1, \mathbf{r}_2 - c\mathbf{s}_2)$
cond	$\mathsf{com} = g^{\mathsf{rsp}} \cdot h^c$	$(\mathsf{com} = \mathbf{A}\mathbf{z}_1 + \mathbf{z}_2 - c\mathbf{t}) \wedge ((\mathbf{z}_i)_i \text{ short})$
Abort?	No	Yes

Dilithium [133]. However, qTESLA [36] provides a textbook example of what can go wrong by trying to apply this idea without checking that the security proof is preserved (in this case, soundness), as it was shown to be completely insecure.

3.2 Beyond Schnorr signatures

For the (vast majority of) problems that do not possess the algebraic properties needed to instantiate Schnorr signatures, there still exist several tricks that enable efficient FS signatures. Scheme designers need to consider two things:

- The soundness error ϵ of the ID protocol is often too large. For example, Stern's code-based protocol has a soundness error $\epsilon = 2/3$. A simple solution is to repeat the protocol k times so that $\epsilon^k \leq 2^{-\lambda}$ for security parameter λ, but finding ways to improve ϵ is also important.
- For some problems, a 3-move ID protocol may be less efficient than an n-move protocol with $n > 3$, or may even not be known.

We first elaborate on the first point. When the soundness ϵ of an ID protocol is too small, the protocol is repeated k times. Typically, all k iterations are performed in parallel (as opposed to sequentially). Parallel repetition is often *expected* by scheme designers to provide exponential soundness ϵ^k, however it is not the case in general; it is proven effective for 3-move *interactive* protocols, but counter-examples exist for protocols with 4 or more moves [20].

Next, we present 3-moves and 5-moves ID schemes. As long as the underlying problem admits some linearity properties, one can build an ID scheme on it [12]. It is the case of all the schemes presented below.
PKP: A 5-move protocol based on the Permuted Kernel Problem (PKP) was proposed in [171], with a soundness error of $\frac{p}{2p-2} \approx 1/2$, where p is the cardinal of the underlying ring. It was later instantiated by PKP-DSS [33].
MQ: The first ID schemes for MQ were proposed by [164]. A key idea of [164] was to use the polar form of $\mathbf{F}$: $\mathbf{G}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{F}(\mathbf{x}_1 + \mathbf{x}_2) - \mathbf{F}(\mathbf{x}_1) - \mathbf{F}(\mathbf{x}_2)$.
$\mathbf{G}$ is bilinear, and this was exploited to propose a 3-move protocol with soundness error 2/3, and a 5-move one with soundness error $1/2 + 1/q \approx 1/2$. The latter protocol was instantiated by MQDSS [49, 165] using the Fiat-Shamir transform.

Codes: Many code-based schemes derive from Stern's elegant protocols [174, 175], which are based on the SD problem. Stern proposed a 3-move with soundness error $2/3$, and a 5-move protocol with soundness error $1/2$. The 3-move version was improved by Veron [180] using the generator matrix of a code instead of its parity check matrix, hence it is often seen as a dual of Stern's protocol. However, most derivatives of Stern's protocol are based on the 5-move variant.
Isogenies: The CSIDH problem has been used to propose an ID scheme that, interestingly, is very similar to the well-known proof of knowledge for graph isomorphism. A useful trick used by SeaSign [65] is to use n public keys; this improves the soundness error down to $\frac{1}{n+1}$. CSI-Fish [34] improved it to $\frac{1}{2n+1}$ by using symmetries specific to isogenies. Both schemes combine this with Merkle trees, which provides a trade-off between signing time and soundness error.
Cut-and-choose: This *generic* technique [124] provides a trade-off between signing time and soundness error. It had been used by [31] to provide MQ-based and PKP-based signatures that are more compact than MQDSS and PKP-DSS.

We end on a note of caution. A recent paper [122] shows that for 5-round ID schemes with k parallel repetitions, the soundness error may be larger than ϵ^k, provides a combinatorial attack against the MQ-based schemes of [49, 165], as well as the PKP-based scheme of [33], and warns that it might apply on 5-round variants of Stern's protocol. Designers of schemes that fit this pattern should be careful.

3.3 Full Domain Hash signatures

Hash-then-sign schemes are among the most intuitive schemes to understand at a high level. The standard way to construct them is via the *Full Domain Hash* (FDH) framework. Let $(\mathsf{sk}, \mathsf{pk})$ be an asymmetric keypair. Associate to it a pair $(f_{\mathsf{pk}}, g_{\mathsf{sk}})$ of efficiently computable functions $f_{\mathsf{pk}} : D \to R$ (surjective) and $g_{\mathsf{sk}} : R \to D$ (injective). We say $(f_{\mathsf{pk}}, g_{\mathsf{sk}})$ is:

- A trapdoor permutation (TP) if:
  (T1) given only pk, f_{pk} is computationally hard to invert.
  (T2) $f_{\mathsf{pk}} \circ g_{\mathsf{sk}}$ is the identity over R.
  (T3) For any y, the distribution of $g_{\mathsf{sk}}(y)$ is (statistically) independent of sk.
  (T4) f_{pk} and g_{sk} are permutations (hence $D = R$).
- A trapdoor preimage sampleable function (TPSF) if it satisfies (T1), (T2), (T3). Hence (T4) is no longer required.
- An average TPSF if it satisfies (T1), (T2), and this relaxation of (T3):
  (T3*) On average over y, the distribution of $g_{\mathsf{sk}}(y)$ is (statistically) independent of sk.

Note that we have the following relation: TP $\Rightarrow$ TPSF $\Rightarrow$ Average TPSF. The FDH framework [21, 22] allows, in its original form, to build hash-then-sign schemes from a hash function and a TP family as in Fig. 4. Note that the function of (3) is a RSA-based TP for whoever knows the factorization $N = p \cdot q$.

Notable efforts at transposing the FDH framework in a post-quantum setting are the code-based schemes CFS [55] and RankSign [88]. The bit-security of sign(msg, sk)

- Compute $H(\mathsf{msg}) = y \in R$;
- Return $\mathsf{sig} \leftarrow f_{\mathsf{sk}}^{-1}(y)$.

verify(msg, pk, sig)

- Accept iff $f_{\mathsf{pk}}(\mathsf{sig}) = H(\mathsf{msg})$.

Figure 4: The Full-Domain Hash (FDH) framework.

CFS scales logarithmically in its parameters, making the scheme impractical, and [82] showed that its security proof requires infeasible parameters. Similarly, [69] showed that RankSign's proposed parameters made the underlying problem easy, and that it required impractical parameters for the scheme to be secure. Both CFS and RankSign indicate that a rigid transposition of FDH framework (using TP) in a post-quantum setting seems highly nontrivial.

Early lattice-based attempts such as GGHSign [95] and NTRUSign [105] instead chose to replace TPs with trapdoor one-way functions (with $|D| \gg |R|$), so that only (T1) and (T2) were verified. In particular, the independence property (T3) was no longer verified. However, (T3) plays an important role in the original security proof of the FDH,[1] which did no longer apply. More critically, each $y \in R$ now admitted many $x_i \in D$ such that $f_{\mathsf{pk}}(x_i) = y$, and the x_i picked by the signing algorithm depended of sk. This dependency was exploited by learning attacks [141, 77] to recover the signing key.

For lattices, the first real progress was done by [93]. Its main contribution was to introduce TPSFs, to prove that they can be used to instantiate the FDH, and to propose provably secure lattice-based TPSFs. Several follow-up schemes have been proposed [137, 76], including Falcon [156].

However, it is not known how to instantiate TPSFs from code-based assumptions. Hence the work of [68, 46] relaxed – again – this notion by proposing average TPSFs, showed that they suffice to instantiate the FDH framework, and proposed a signature scheme based on code-based average TPSFs, Wave [68]. Interestingly, this idea was proposed independently by [50], which show that lattice-based average TPSFs require milder parameters than TPSFs, hence improving upon the efficiency of some TPSF-based lattice signatures [29].

$$\text{Trapdoor Permutation} \Rightarrow \text{TPSF} \Rightarrow \text{Average TPSF}.$$

Multivariate cryptography encountered and solved this problem independently. It was first noticed in [163] that some multivariate hash-then-sign schemes relied on a trapdoor function that only verified (T1) and (T2). Hence [163] introduced of a salt during the signing procedure in order to satisfy (T3) and enable a FDH-style proof. This solution is now used by GeMSS [43] and Rainbow [73].

[1] In the case of TPs, the situation is simpler since (T2) + (T4) ⇒ (T3).

3.4 Diffie-Hellman and El Gamal

The Diffie-Hellman (DH) key-exchange protocol [72], as well as the encryption scheme by El Gamal that is derived from it [78], are staples of classical public key cryptography. El Gamal has been notably easier to adapt to PQ assumptions than DH. Classically, DH relies on (2), which provides a simple way for two parties to agree on a shared secret g^{ab}, by instantiating Fig. 5 with Table 3 (column 2). Unfortunately, such a simple relation is harder to obtain with PQ assumptions, as we will see.

Isogenies over elliptic curves are the most natural candidate to instantiate Fig. 5. Unfortunately, the most natural way to do that requires either ordinary curves [57, 162] – which parameters don't scale well [66] –, or supersingular curves with a restricted class of isogenies like CSIDH [45] – which quantum security is debated [37, 147]. A "standard" approach is to use supersingular curves with low-degree isogenies, however it requires to apply the private isogeny $\phi_A : E \to E_A$ to two special points P_B, Q_B of the elliptic curve E, and send the result in addition to E_A. Only with this extra information can the two parties agree on a common curve E_{AB}. A straightforward adaptation of DH to codes and lattices is challenging as well, this time due to *noise*. For example, a rigid transposition with LWE gives:

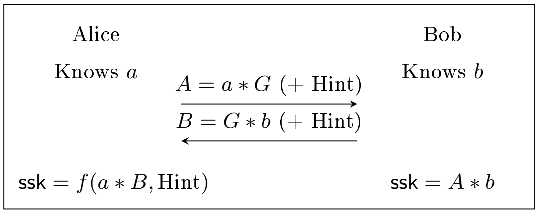


Figure 5: Diffie-Hellman with Reconciliation.

$$(\mathbf{s}_a^t \cdot \mathbf{A} + \mathbf{e}_a^t)\mathbf{s}_b \approx \mathbf{s}_a^t(\mathbf{A} \cdot \mathbf{s}_b + \mathbf{e}_b) \tag{6}$$

Both parties would end up with "noisy secrets" that differ on their lower bits, which is problematic. In a purely non-interactive setting, this approach does not seem to work, except if q is very large, say $q \geq 2^{\lambda}$, which is impractical [102]. This is resolved in [74, 149] by sending a hint indicating "how to round the noisy secret". Note that this approach comes at the cost of non-interactivity.

Table 3 summarizes the two approaches to achieve "post-quantum DH" (besides CSIDH). In addition to being interactive, these solutions cannot be used with static key shares, as it would enable key-recovery attacks [86, 90]. As such, they cannot be used as drop-in replacements to non-interactive (semi-)static DH.

Many desirable properties of classical DH are lost in translation when transposing it to a PQ setting. As such, most practical schemes take El Gamal as a starting point instead, replacing DLOG with LWE [140, 170], Learning With Rounding (LWR) [61], or SIDH [117]. Schemes that rely on "trapdoors" – like

Table 3: A few ways to instantiate Fig. 5.

	(EC)DH	SIDH [118, 84]	LWE [74, 149]
G	$g \in \mathbb{G}$	$(P_i, Q_i)_i$	$\mathbf{A} \in R_q^{k \times k}$
a	$a \in \lvert\mathbb{G}\rvert$	Isogeny $\phi_A : E \to E_A$	$(\mathbf{s}_a, \mathbf{e}_a)$ short
A	g^a	$E_A, \phi_A(P_B), \phi_A(Q_B)$	$\mathbf{s}_a^t \cdot \mathbf{A} + \mathbf{e}_a^t$
B	g^b	$E_B, \phi_B(P_A), \phi_B(Q_A)$	$\mathbf{A} \cdot \mathbf{s}_b + \mathbf{e}_b$
Hint	No	Two-way	One-way
Static	Yes	No	No

McEliece [135, 24] or BIKE-2 [9] – are more akin to RSA encryption, though this analogy is a weaker one.

4 Return of Symmetric Cryptography

Another takeaway is that, despite PQC being mostly a public-key matter, symmetric cryptography plays a surprisingly important role and should not be neglected. In particular, two families of signatures based on one-way and hash functions have emerged, with two radically different philosophies:

- Hash-based signatures treat hash functions as *black boxes* and build signatures using only generic data structures and combinatorial tricks, see Sec. 4.1.
- Signatures based on zero-knowledge proofs treat one-way functions as *white boxes* and leverage knowledge of their internal structure to maximize their efficiency, see Sec. 4.2.

Interestingly, some techniques developed by these schemes have also benefited more "standard" schemes. Examples are Merkle trees, used by multivariate [35] and isogeny-based [65, 34] schemes, or the *cut-and-choose* technique [124].

4.1 Hash-based signatures

Hash-based signatures (HBS) are a peculiar family of schemes for two reasons; (a) they rely solely on the hardness properties of hash functions, (b) they follow a paradigm of their own. At a high level:

- The public key pk commits secret values using one or more hash functions.
- Each signature reveals (intermediate) secret values that allow to recompute pk and convince the verifier that the signer does indeed know sk.

Lamport's HBS [129] epitomizes this idea. In its simplest form, the public key is: $\mathsf{pk} = (\mathsf{pk}_{i,0}, \mathsf{pk}_{i,1})_{i \in [\lambda]} = (H(\mathsf{sk}_{i,0}), H(\mathsf{sk}_{i,1}))_{i \in [\lambda]}$, and the signature of a message $\mathsf{msg} = (b_i)_i \in \{0,1\}^\lambda$ is $\mathsf{sig} = (\mathsf{sk}_{i,b_i})_i$. The verifier can then hash sig componentwise and check it against pk. It is easily shown that Lamport's signature scheme is secure under the preimage resistance of H. However, there are two caveats:

- pk and sig require $O(\lambda^2)$ bits, which is rather large.

- It is a one-time signature (OTS), meaning it is only secure as long as it performs no more than one signature.

For four decades, several tricks have been proposed to mitigate these caveats. Because of the unstructured nature of hash functions, these tricks typically rely on combinatorics and/or generic data structures.

One line of research proposes efficient data structures that use OTS as building blocks. By hashing public keys into a tree, Merkle trees [136] allow to improve efficiency and sign more than one message. Goldreich trees [94] use trees' leaves to sign other trees' roots. Both ideas can be combined, as done by SPHINCS($^+$) [26, 27, 115]. Finally, efficient Merkle tree traversal algorithms were proposed [176].

Another line of research proposed more efficient OTS. The most efficient one so far is a variant of Winternitz's OTS (see [136, 42]), called WOTS+ [114], which uses bitmasks to rely on second-preimage resistance – instead of collision resistance for the original scheme. Stateless few-time signatures (FTS) were also proposed, such as BiBa [151], HORS (Hash to Obtain Random Subsets) [159], a HORS variant with trees, HORST [26], one with PRNGs, PORS [11], and another one with forests, FORS [27, 115]. These can be used to build *stateless* signatures, discussed below.

These tools allow to build hash-based signatures, which can be categorized in two families: *stateful* and *stateless* signatures.

Stateful schemes require the signer to maintain an internal state in order to keep track of the key material used. This encompasses XMSS, its multi-tree variant XMSS$^{\mathrm{MT}}$ and LMS, all recently standardized by NIST [52]. Stateful schemes can be efficient but their statefulness is often an undesirable property.

Stateless signatures set their parameters so that, even without maintaining a state, signing many messages will preserve security with overwhelming probability. As a result, they are less efficient than their stateful counterparts, but more flexible. For example, SPHINCS$^+$ [27, 115] combines Merkle and Goldreich trees with WOTS+ as an OTS, FORS as a FTS, plus a few other tricks.

4.2 Signatures based on ZKPs and OWFs

Signatures based on zero-knowledge proofs (ZKPs) and one-way functions (OWFs) leverage this principle:

- The public key is $\mathsf{pk} = F(\mathsf{sk})$, where F is a OWF.
- A signature is a ZKP that $\mathsf{pk} = F(\mathsf{sk})$; using the MPC-in-the-head [116].

Note that all Fiat-Shamir signatures can already be interpreted as ZKP that $\mathsf{pk} = F(\mathsf{sk})$, however they usually leverage algebraic structure to gain efficiency, and as a result rely on assumptions that are algebraic in nature.

The protocols discussed here are fully generic as they work with any OWF. This is done by leveraging the *MPC-in-the-head* technique [116]. This technique creates non-interactive proofs for an arbitrary circuit (Boolean or arithmetic), by simulating the execution of an MPC (*multiparty computation*) protocol, committing to the execution, and revealing the state of a subset of the parties in order to let the verifier (partially) check correctness of the execution. Two parallel yet connected lines of research turned this abstract idea into a reality.

The first line of research provides protocols for generic statements. Such protocols have only recently become practical, see ZKB++[48] and KKW [124]. For bit-security λ and a circuit with $|C|$ AND gates, total proof sizes are $O(\lambda|C|)$, for ZKB++, and $O(\lambda|C|/\log n)$, for KKW, respectively, where the *cut-and-choose* approach of KKW allows a trade-off between signing and signature size, via the parameter n. For boolean (resp. arithmetic) circuits of cryptographic sizes, these two schemes (resp. the sacrificing method [17]) are the current state of the art.

The second line of research provides circuits with low multiplicative complexity. Because of their unusual constraints, their internal structure is typically very different from classical symmetric primitives and they require new approaches to be studied. Prominent examples are LowMC [8], which has been extensively studied [75, 119, 131], or the Legendre PRF [59, 96]. Note that these primitives have applications that go far beyond PQC; for example, the Legendre PRF is used by the Ethereum 2.0 protocol.

Combining these two lines of research, one obtain signature schemes. For example, Picnic [183] combines LowMC with either ZKB++ or KKW, BBQ [67] combines AES with KKW, and finally LegRoast [32] combines the Legendre PRF with the sacrificing method [17]. Due to the novely of this approach, it is likely that we will see many more schemes based on it in the future.

5 The Implementation Challenges in PQC

This section discusses the implementation challenges in PQC; specifically discussing attacks via implementation pitfalls and side-channels, countermeasures, and finally the jungle of embedded devices and use-cases for PQC schemes. We somewhat focus on NIST PQC candidates due to similarities in the operations each PQC family requires.

5.1 Decryption Failures and Reaction Attacks

Attacks based on decryption failures – also known as reaction attacks – were first discovered about 20 years ago, with an attack [103] on the McEliece [135] and Ajtai-Dwork [3] cryptosystems, and another [112] on NTRU [107]. They were forgotten for more than a decade before being recently rediscovered. It is clear by now that designers of noisy cryptosystems, such as lattice-based and code-based, need to take these into account. We explain how reaction attacks work and how to thwart them. At a high level, *all* lattice-based and code-based encryption schemes follow this high-level description: $\mathsf{ct} = \mathsf{pk} \cdot \mathbf{e} + \mathbf{e}' + \mathsf{Encode}(\mathsf{msg})$, where $\mathsf{Encode}(\mathsf{msg})$ is an encoding of msg and $(\mathbf{e}, \mathbf{e}')$ is a noisy error vector. The decryption key sk is used to obtain $\mathsf{Encode}(\mathsf{msg})$ plus some noise, then recover msg. However, this may fail for a small portion of the admissible $(\mathbf{e}, \mathbf{e}')$, and this portion depends on sk. The high-level strategy of reaction attacks uses:

- **Precomputation.** Precompute "toxic" errors $(\mathbf{e}, \mathbf{e}')$ that have a high probability of leading to decryption failures;
- **Query.** Use these toxic errors to send ciphertexts to the target; observe decryption failures.
- **Reconstruction.** Deduce sk from the decryption failures.

Note that reaction attacks are CCA attacks. In CCA schemes, $(\mathbf{e}, \mathbf{e}')$ is generated by passing msg and/or pk into a pseudo-random generator (PRG), so adversaries have to find toxic vectors through exhaustive search. Hence precomputation is often the most computationally intensive phase.

Reaction attacks have been proposed against code-based schemes in the Hamming metric [100], in the rank metric [166], and for lattice-based schemes [60, 64, 101]. Interestingly, attacks against schemes that use lattices or the Hamming metric are very geometric (learning the geometry of the private key), whereas those that target rank metric schemes learn algebraic relations.

For lattice-based schemes, *directional failure boosting* [62] allows, once a toxic error $(\mathbf{e}, \mathbf{e}')$ has been found, to find many more at little cost. Therefore, lattice schemes *must* keep their failure probability negligible, as they are otherwise directly vulnerable to reaction attacks. No such conclusion has been made for code-based schemes yet, but we recommend scheme designers to err on the safe side. Scheme designers need to consider two things with respect to reaction attacks. First, the probability of decryption failures should be negligible.

- This can be achieved by selecting the parameters accordingly, as done by Kyber [170], Saber [61], HQC [1] and FrodoKEM [140]. One may even eliminate them completely like NTRU [185] and NTRU Prime [25], but this may result in slightly larger parameters.
- Another solution is to use redundancy; KEMs need to encapsulate a symmetric key of λ bits, however schemes can often encrypt a much larger message msg. One can use the extra bits to embed an error-correcting code (ECC). However, this solution has two caveats. First, the ECC should be constant-time (e.g., XEf [185] and Melas codes [104]), as timing attacks have been observed when that was not the case [63]. Second, this requires to perform a tedious analysis of the noise distribution; incorrect analyses have led to theoretical attacks [64, 101].

Second, schemes with decryption failures – even negligible – should use CCA transforms that take these into account. In effect, most PQ KEMs in this situation use variants of the transforms described [108], which do handle them.

5.2 Implementation Attacks in PQC

Before NIST began their PQC standardization effort, many PQC schemes were susceptible to implementation attacks; meaning that due to bad coding practices, some attack vectors were found which led to successful attacks. Definition 5 in [111] provides a fairly formal definition for isochronous algorithms (i.e., an algorithm with no timing leakage) which allows us to differentiate between these initial implementation attacks, of which many did not qualify. Good programming practices exist for ensuring timing analysis resilience and have been well discussed before[2]. These practices cover much more low-level instances of isochronous designs; as conditional jumps, data-dependent branching, and memory accesses of secret information can also lead to detrimental attacks. Some tools such as `ctgrind`, `ctverif`, and `flow-tracker` exist to check whether functions are isochronous, however with operations in PQC such as rejection sampling it is not clear how effective these tools will be. Thus, it would also be prudent to check post-compilation code of the sensitive operations within an implementation.

The first types of implementation attacks on PQC were mainly on the BLISS signature scheme and exploited the cache-timing leakages from the Gaussian samplers, as they mostly operate by accessing pre-computed values stored in memory [40, 152]. The attacks use the FLUSH+RELOAD [182] technique and exploit cache access patterns in the samplers to gain access to some coefficients of values that are added during the signature's calculation. However, optimisations to the Gaussian samplers, such as using guide-tables, and non-isochronous table access enabled these attacks. More leakage sources and implementation attacks against the StrongSwan implementation of BLISS were also found [79], which range from data dependent branches present in the Gaussian sampling algorithm to using branch tracing in the signature's rejection step. These attacks can be mitigated by bypassing conditional branches; that is, using a consistent access pattern (e.g., using linear searching of the table) and having isochronous runtime. In particular, making Gaussian samplers provably secure and statistically proficient have been researched [111] and thus should be followed for secure implementations of lattice-based schemes such as Falcon and FrodoKEM or more advanced primitives such as IBE and FHE.

Although these attacks are on a scheme's implementation, rather than something inherently insecure in its algorithm, they have acted as a cautionary note for how some schemes have operations, which do not use secret information, but could be described as *sensitive* as if they are implemented incorrectly, they can lead to a successful attack. A clear example of this is for Gaussian samplers, which is why they were not used in Dilithium. Once an attacker finds the error vector, $\mathbf{e}$, using these side-channels from a LWE equation of the form $\mathbf{b} = \mathbf{A} \times \mathbf{s} + \mathbf{e} \mod q$, then gaining the secret can be achieved using Gaussian elimination. Moreover, it is not always necessary to find the entire secret, as was the case in the past for RSA [53], and side-channels can be combined with lattice reduction algorithms efficiently to significantly improve attacks on post-quantum schemes. This has been built into a framework [58], which builds in side information into lattice reduction algorithms in order to predict the performance of lattice attacks and estimate the security loss for given side-channel information.

Another sensitive component is in the transient version of the HQC cryptosuite proposed during the NIST PQC standardization process. In the proposed (but now deprecated) reference implementation of decryption, the most costly component was a multiplication in $\mathbb{F}_2[X]/(X^n - 1)$. The crucial operation dur-

[2] See for example https://www.bearssl.org/constanttime.html.

ing decryption is a sparse-dense polynomial multiplication over $\mathbb{F}_2[X]$. At one point in time (specifically, for less than a month in the overall NIST PQC process), it was proposed to use an special algorithm for sparse-dense multiplication, where the complexity of the multiplication was better than the obvious schoolbook algorithm, by utilizing the sparseness of the secret-key polynomial. That is, the multiplication would only access the secret-key polynomial h times, for a secret-key containing only h 1's. In particular, a further, "shielded" version of this algorithm was proposed which applied a permutation (on the memory-access locations) in order to attempt to hide the fact that only h locations were ever accessed in the memory cells corresponding to the secret-key polynomial, while retaining the efficiency benefits of an algorithm specialized to the case of sparse-dense polynomial multiplication. Unfortunately, if an adversary can only observe the memory cells accessed during memory (even without seeing the contents of those memory cells), then – by analogy to an "Oblivious RAM" adversary, the secret key can be directly recovered after one decryption is performed.

A sensitive component that can potentially affect all PQC candidates is in the Fujisaki-Okamoto (FO) transformation. This component is required in lattice-based and code-based KEMs in order to covert the CPA-secure part into an IND-CCA secure scheme. However, it has been shown that this operation is also sensitive to timing attacks, even though the operations do not use any secret information. This attack [99] was shown on FrodoKEM, and was enabled due to its use of non-isochronous `memcmp` in the implementation of the ciphertext comparison step, which allows recovery of the secret key with about 2^{30} decapsulation calls. This attack is directly applied to FrodoKEM, but is likely that other PQC candidates such as BIKE, HQC, and SIKE are also susceptible.

An algorithm used within the FO transform is Keccak, or more specifically SHAKE, which was standardized by NIST in FIPS-202 for SHA-3 and is used extensively within NIST PQC candidates for so-called seed-expansion and computation of the shared secret. This symmetric operation is also sensitive to side-channels and could potentially lead to recovery of the shared-secret generated in the KEM. In particular, a single trace attack was demonstrated on the Keccak permutation in the ephemeral key setting [123], but seemingly realistic only on 8-bit devices.

Finally we consider the peculiar nature of BIKE's (sensitive) decryption module. The BIKE decryption algorithm is naturally designed to proceed in a repetitive sequence of steps. Some operations are performed, then the message is properly decrypted, or not. Such operations can then be repeated, and the likelihood of proper decryption will increase. Unlike most other PQ decryption procedures, the BIKE decryption algorithm is not inherently isochronous, nor is the decryption failure rate well-understood. Given the real-world requirement that all secret-sensitive procedures are isochronous, it has been proposed to therefore artificially truncate this iterative decryption procedure at some fixed number of steps. Experimentally, a round-count as small as 10 is sufficient to guarantee proper decryption. However, in contrast to the case of lattice-based KEMs, there is no mathematical guarantee that, e.g., 10 iterations is sufficient to reduce the decryption failure rate of the scheme below 2^{λ}, where $\lambda \in \{128, 192, 256\}$ is the concrete security parameter.[3] Therefore, despite the BIKE scheme being designed as first a CPA scheme along with a CPA-to-CCA transform implemented as low cost, the BIKE team has only formally claimed CPA-security (that is, ephemeral key security) for their construction, as opposed to CCA-security (that is, long-term key security). It remains open to provide a "proper analysis" of the BIKE decryption algorithm guaranteeing sufficient precision of decryption failures to ensure long-term key security for the scheme.

5.3 Side-Channels and Countermeasures

In the Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process [4] it is stated that:

> *NIST hopes to see more and better data for performance in the third round. This performance data will hopefully include implementations that protect against side-channel attacks, such as timing attacks, power monitoring attacks, fault attacks, etc.*

In their initial submission requirements [142] NIST also noted that "schemes that can be made resistant to side-channel attacks at minimal cost are more desirable than those whose performance is severely hampered by any attempt to resist side-channel attacks". Thus, some of the remaining candidates also have offer masked implementations, or this has been contributed by the research community.

Migliore et al. [138] demonstrate DPA weaknesses in the unmasked Dilithium implementation, and in addition to this provide a masking scheme using the Ishai-Sahai-Wagner (ISW) probing model following the previous techniques for masking GLP and BLISS [15, 16]. Like the previous provably secure masking schemes, they alter some of the procedures in Dilithium by adding in efficient masking of its sensitive operations. Moreover, some parameters are changed to gain extra performance efficiencies in the masked design, such as making the prime modulus a power-of-two, which increases the performance by 7.3 to 9 times compared to using the original prime modulus during masking. A power-of-two modulus means the optimised multiplication technique, the NTT multiplier, is no longer possible so they proposed Karatsuba multiplication. The results for key generation and signing are between 8 to 12 times slower for order 2 masking and 13 to 28 times slower for order 3 masking, compared to the reference implementations. This is also backed-up by experimental leakage tests on the masked designs.

Similarly, Verhulst [179] provides DPA on Saber, as well as developing a masking scheme for its decryption protocol, which is later extended in [19]. The masking schemes only use additive first-order masking which thus makes it only 2 to 2.5 times slower than being unprotected. However it is probably still vulnerable to template attacks [143]. Saber lends itself to practical masking due to its use of LWR, as opposed to other KEMs using (M-)LWE. However, Saber uses

[3] Known, formal analyses guarantees are closer to 2^{-40} at 128-bit security.

a less efficient multiplication method (a combination of Toom-Cook, Karatsuba, and schoolbook multiplication) compared to schemes which use number theoretic transform (NTT); thus it is an interesting open question as to whether NTT is the most practical multiplication method (due to its conflict with efficient masking) and how these masked PQC schemes practically compare, particularly with the recent research improving the performance of Saber and others using NTTs [51].

NTRU and NTRU Prime both have the potential of using a combination of Toom-Cook and Karatsuba to speed-up their polynomial multiplication, thus whether they can reuse techniques from Saber's masked implementation is an important research question. NTRU Prime in particular requires masking since some power analysis attacks can read off the secret key with the naked eye [113]. Attacks on these multiplication methods, which are in the time-domain, are likely to be simpler than those in the NTT or FFT domains as there is only one multiplication per coefficient of the secret, which thus makes protection of this multipliers more urgent. A single-trace power analysis attack on FrodoKEM exploits the fact that the secret matrix is used multiple times during the matrix multiplication operation, enabling horizontal differential power analysis [38].

Correlation power analysis and algebraic key recovery attacks have also been shown on the schemes Rainbow and UOV [144] by targeting the secret maps within the MQ signature schemes, during the matrix-vector computations. This attack is relevant for many MQ schemes that use the affine-substitution quadratic-affine (ASA) structure. They also discuss countermeasures to simple and differential power analysis by using standard methods seen before such as shuffling of the indices or adding a pseudo-random matrix (i.e., additive masking).

QcBits, a variant of McEliece PKE, was shown to be susceptible to DPA [161]. The attack partially recovers the secret key during the syndrome computation of the decoding phase. They also propose a simple countermeasure for the syndrome calculation stage, which exploits the fact that since QC-MDPC (quasi-cyclic moderate-density parity-check) codes are linear, the XOR of two codewords is another codeword. Thus, a codeword can be masked by XORing it with another random codeword before the syndrome calculation.

This attack was then extended [173] to recover the *full* secret of QcBits, with more accuracy, using a multi-trace attack. Moreover, using the DPA countermeasures proposed in [161] and in the ephemeral key setting, they provide a single-trace attack on QcBits. Lastly and most interestingly, they describe how these attacks can be applied to BIKE, by targetting the private syndrome decoding computation stage where long-term keys are utilized. For ephemeral keys, the multi-target attacks are not applicable, however the single-trace attack can be applied to recover the private key and also the secret message.

Classic McEliece is also not immune from side-channel attacks targeting this operation. A reaction attack [128] using iterative chunking and information set decoding can enable recovery of the values of the error vector using a single decryption oracle request.

Masking schemes which use matrix multiplication have the potential to be efficiently masked using affine masking (i.e., a combination of additive and multiplicative masking) similarly used in the Advanced Encryption Standard (AES) [87]. First-order additive masking has already been proposed for FrodoKEM [109]. Warnings for side-channel protection were also seen in Picnic, where the attack was able to recover the shared secret and the secret key, by targetting the LowMC block cipher, a core component to the signature scheme [91].

PQC schemes have also been shown to be susceptible to cold-boot attacks [154, 6], which was previously shown on NTRU [145]. Cold-boot attacks exploit the fact that secret data can remain in a computer's memory (DRAM) after it is powered down and supposedly deleted. Albrecht et al. [6] describe how to achieve this by attacking the secret-keys stored for use in the NTT multiplier in Kyber and NewHope, and after some post-processing using lattice reductions, is able to retrieve the secret-key.

Fault attacks have also been investigated for PQC schemes. One of the most famous (microarchitectural) fault attacks is the Rowhammer exploit (CVE-2015-0565), which allows unprivileged attackers to corrupt or change data stored in certain, vulnerable memory chips, and has been extended to other exploits such as RAMBleed (CVE-2019-0174). QuantumHammer [139] utilises this exploit to recover secret key bits on LUOV, a second round NIST PQC candidate for multivariate-quadratic signatures. The attack does somewhat exploit the 'lifted' algebraic structure that is present in LUOV, so whether this attack could be applied to other PQC schemes is an open question.

Determinism in signatures is generally considered preferable from a security perspective, as attacks are possible on randomly generated nonces (e.g., [80]). This prompted EdDSA, which uses deterministically generated nonces. NIST [4] noted the potential for nonce reuse in PQC schemes such as Kyber. Indeed, fault attacks which exploit the scheme's determinism have been demonstrated on SPHINCS$^+$ [44] and Dilithium [41, 157], with EdDSA also showing susceptibility to DPA [167]. As such, some PQC candidates offer an optional non-deterministic variant, such as SPHINCS$^+$ using `OptRand`, or random *salt* used in Dilithium, Falcon, GeMSS, Picnic, and Rainbow.

An interesting alternative to mitigating these fault attacks (and randomness failures) is by using *hedging*, which creates a middle-ground between fully deterministic and fully probabilistic signatures, by deriving the per-signature randomness from a combination of the secret-key, message, and a nonce. This is formalized for Fiat-Shamir signatures and apply the results to hedged versions of XEdDSA, a variant of EdDSA used in the Signal messaging protocol, and to Picnic2, and show hedging mitigates many of the possible fault attacks [10].

Key reuse attacks, which have been shown to cause issues for real-world implementations of the EMV ("Europay, Mastercard, Visa") standard [70], are also applicable in PQC; such as lattice-based schemes [86], supersingular isogeny-based schemes [90], and potentially more.

We continue the practical discussions on PQC in the full version of this paper [110], focusing on embedded implementations and use cases, and then providing an overview of how PQC is being standardized, what new protocols are being designed, and any large scale experiments that have been conducted thus far.

References

[1] C. Aguilar Melchor, N. Aragon, S. Bettaieb, L. Bidoux, O. Blazy, J.-C. Deneuville, P. Gaborit, E. Persichetti, and G. Zémor. *HQC.* Tech. rep. National Institute of Standards and Technology, 2019.

[2] M. Ajtai. "Generating Hard Instances of Lattice Problems (Extended Abstract)". In: *28th ACM STOC.* 1996.

[3] M. Ajtai and C. Dwork. "A Public-Key Cryptosystem with Worst-Case/Average-Case Equivalence". In: *29th ACM STOC.* 1997.

[4] G. Alagic, J. Alperin-Sheriff, D. Apon, D. Cooper, Q. Dang, J. Kelsey, Y.-K. Liu, C. Miller, D. Moody, R. Peralta, et al. "Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process". In: *NIST, Tech. Rep., July* (2020).

[5] M. R. Albrecht, B. R. Curtis, A. Deo, A. Davidson, R. Player, E. W. Postlethwaite, F. Virdia, and T. Wunderer. "Estimate All the LWE, NTRU Schemes!" In: *SCN 18.* 2018.

[6] M. R. Albrecht, A. Deo, and K. G. Paterson. "Cold Boot Attacks on Ring and Module LWE Keys Under the NTT". In: *IACR TCHES* 3 (2018).

[7] M. R. Albrecht, R. Player, and S. Scott. "On the concrete hardness of Learning with Errors". In: *J. Math. Cryptol.* 3 (2015).

[8] M. R. Albrecht, C. Rechberger, T. Schneider, T. Tiessen, and M. Zohner. "Ciphers for MPC and FHE". In: *EUROCRYPT.* 2015.

[9] N. Aragon, P. Barreto, S. Bettaieb, L. Bidoux, O. Blazy, J.-C. Deneuville, P. Gaborit, S. Gueron, T. Guneysu, C. Aguilar Melchor, R. Misoczki, E. Persichetti, N. Sendrier, J.-P. Tillich, G. Zémor, and V. Vasseur. *BIKE.* Tech. rep. National Institute of Standards and Technology, 2019.

[10] D. F. Aranha, C. Orlandi, A. Takahashi, and G. Zaverucha. "Security of Hedged Fiat-Shamir Signatures Under Fault Attacks". In: *EUROCRYPT.* 2020.

[11] J.-P. Aumasson and G. Endignoux. "Improving Stateless Hash-Based Signatures". In: *CT-RSA.* 2018.

[12] M. Backendal, M. Bellare, J. Sorrell, and J. Sun. "The Fiat-Shamir Zoo: Relating the Security of Different Signature Variants". In: *NordSec.* 2018.

[13] S. Bai and S. D. Galbraith. "An Improved Compression Technique for Signatures Based on Learning with Errors". In: *CT-RSA.* 2014.

[14] B. Barak and M. Mahmoody-Ghidary. "Merkle's Key Agreement Protocol is Optimal: An $O(n^2)$ Attack on Any Key Agreement from Random Oracles". In: *Journal of Cryptology* 3 (2017).

[15] G. Barthe, S. Belaïd, T. Espitau, P.-A. Fouque, B. Grégoire, M. Rossi, and M. Tibouchi. "Masking the GLP Lattice-Based Signature Scheme at Any Order". In: *EUROCRYPT.* 2018.

[16] G. Barthe, S. Belaïd, T. Espitau, P.-A. Fouque, M. Rossi, and M. Tibouchi. "GALACTICS: Gaussian Sampling for Lattice-Based Constant-Time Implementation of Cryptographic Signatures, Revisited". In: *ACM CCS.* 2019.

[17] C. Baum and A. Nof. "Concretely-Efficient Zero-Knowledge Arguments for Arithmetic Circuits and Their Application to Lattice-Based Cryptography". In: *PKC.* 2020.

[18] A. Becker, L. Ducas, N. Gama, and T. Laarhoven. "New directions in nearest neighbor searching with applications to lattice sieving". In: *SODA*. 2016.

[19] M. V. Beirendonck, J.-P. D'Anvers, A. Karmakar, J. Balasch, and I. Verbauwhede. *A Side-Channel Resistant Implementation of SABER*. Cryptology ePrint Archive, Report 2020/733. 2020.

[20] M. Bellare, R. Impagliazzo, and M. Naor. "Does Parallel Repetition Lower the Error in Computationally Sound Protocols?" In: *38th FOCS*. 1997.

[21] M. Bellare and P. Rogaway. "Random Oracles are Practical: A Paradigm for Designing Efficient Protocols". In: *ACM CCS 93*. 1993.

[22] M. Bellare and P. Rogaway. "The Exact Security of Digital Signatures: How to Sign with RSA and Rabin". In: *EUROCRYPT'96*. 1996.

[23] *Post-Quantum Cryptography*. 2009.

[24] D. J. Bernstein, T. Chou, T. Lange, I. von Maurich, R. Misoczki, R. Niederhagen, E. Persichetti, C. Peters, P. Schwabe, N. Sendrier, J. Szefer, and W. Wang. *Classic McEliece*. Tech. rep. National Institute of Standards and Technology, 2019.

[25] D. J. Bernstein, C. Chuengsatiansup, T. Lange, and C. van Vredendaal. *NTRU Prime*. Tech. rep. National Institute of Standards and Technology, 2019.

[26] D. J. Bernstein, D. Hopwood, A. Hülsing, T. Lange, R. Niederhagen, L. Papachristodoulou, M. Schneider, P. Schwabe, and Z. Wilcox-O'Hearn. "SPHINCS: Practical Stateless Hash-Based Signatures". In: *EUROCRYPT*. 2015.

[27] D. J. Bernstein, A. Hülsing, S. Kölbl, R. Niederhagen, J. Rijneveld, and P. Schwabe. "The SPHINCS$^+$ Signature Framework". In: *ACM CCS*. 2019.

[28] D. J. Bernstein and B.-Y. Yang. *Asymptotically faster quantum algorithms to solve multivariate quadratic equations*. Cryptology ePrint Archive, Report 2017/1206. 2017.

[29] P. Bert, P.-A. Fouque, A. Roux-Langlois, and M. Sabt. "Practical Implementation of Ring-SIS/LWE Based Signature and IBE". In: *Post-Quantum Cryptography - 9th International Conference, PQCrypto*. 2018.

[30] L. Bettale, J. Faugère, and L. Perret. "Solving polynomial systems over finite fields: improved analysis of the hybrid approach". In: *ISSAC*. 2012.

[31] W. Beullens. "Sigma Protocols for MQ, PKP and SIS, and Fishy Signature Schemes". In: *EUROCRYPT*. 2020.

[32] W. Beullens and C. de Saint Guilhem. "LegRoast: Efficient Post-quantum Signatures from the Legendre PRF". In: *Post-Quantum Cryptography - 11th International Conference, PQCrypto*. 2020.

[33] W. Beullens, J.-C. Faugère, E. Koussa, G. Macario-Rat, J. Patarin, and L. Perret. "PKP-Based Signature Scheme". In: *INDOCRYPT*. 2019.

[34] W. Beullens, T. Kleinjung, and F. Vercauteren. "CSI-FiSh: Efficient Isogeny Based Signatures Through Class Group Computations". In: *ASIACRYPT*. 2019.

[35] W. Beullens, B. Preneel, and A. Szepieniec. "Public Key Compression for Constrained Linear Signature Schemes". In: *SAC*. 2019.

[36] N. Bindel, S. Akleylek, E. Alkim, P. S. L. M. Barreto, J. Buchmann, E. Eaton, G. Gutoski, J. Kramer, P. Longa, H. Polat, J. E. Ricardini, and G. Zanon. *qTESLA*. Tech. rep. National Institute of Standards and Technology, 2019.

[37] X. Bonnetain and A. Schrottenloher. "Quantum Security Analysis of CSIDH". In: *EUROCRYPT*. 2020.

[38] J. W. Bos, S. Friedberger, M. Martinoli, E. Oswald, and M. Stam. "Assessing the Feasibility of Single Trace Power Analysis of Frodo". In: *SAC*. 2019.

[39] L. Both and A. May. "Decoding Linear Codes with High Error Rate and Its Impact for LPN Security". In: *Post-Quantum Cryptography - 9th International Conference, PQCrypto*. 2018.

[40] L. G. Bruinderink, A. Hülsing, T. Lange, and Y. Yarom. "Flush, Gauss, and Reload - A Cache Attack on the BLISS Lattice-Based Signature Scheme". In: *CHES*. 2016.

[41] L. G. Bruinderink and P. Pessl. "Differential Fault Attacks on Deterministic Lattice Signatures". In: *IACR TCHES* 3 (2018).

[42] J. Buchmann, E. Dahmen, S. Ereth, A. Hülsing, and M. Rückert. "On the Security of the Winternitz One-Time Signature Scheme". In: *AFRICACRYPT 11*. 2011.

[43] A. Casanova, J.-C. Faugère, G. Macario-Rat, J. Patarin, L. Perret, and J. Ryckeghem. *GeMSS*. Tech. rep. National Institute of Standards and Technology, 2019.

[44] L. Castelnovi, A. Martinelli, and T. Prest. "Grafting Trees: A Fault Attack Against the SPHINCS Framework". In: *Post-Quantum Cryptography - 9th International Conference, PQCrypto*. 2018.

[45] W. Castryck, T. Lange, C. Martindale, L. Panny, and J. Renes. "CSIDH: An Efficient Post-Quantum Commutative Group Action". In: *ASIACRYPT*. 2018.

[46] A. Chailloux and T. Debris-Alazard. "Tight and Optimal Reductions for Signatures Based on Average Trapdoor Preimage Sampleable Functions and Applications to Code-Based Signatures". In: *PKC*. 2020.

[47] A. Chailloux, M. Naya-Plasencia, and A. Schrottenloher. "An Efficient Quantum Collision Search Algorithm and Implications on Symmetric Cryptography". In: *ASIACRYPT*. 2017.

[48] M. Chase, D. Derler, S. Goldfeder, C. Orlandi, S. Ramacher, C. Rechberger, D. Slamanig, and G. Zaverucha. "Post-Quantum Zero-Knowledge and Signatures from Symmetric-Key Primitives". In: *ACM CCS*. 2017.

[49] M.-S. Chen, A. Hülsing, J. Rijneveld, S. Samardjiska, and P. Schwabe. "From 5-Pass MQ-Based Identification to MQ-Based Signatures". In: *ASIACRYPT*. 2016.

[50] Y. Chen, N. Genise, and P. Mukherjee. "Approximate Trapdoors for Lattices and Smaller Hash-and-Sign Signatures". In: *ASIACRYPT*. 2019.

[51] C.-M. M. Chung, V. Hwang, M. J. Kannwischer, G. Seiler, C.-J. Shih, and B.-Y. Yang. *NTT Multiplication for NTT-unfriendly Rings*. Cryptology ePrint Archive, Report 2020/1397. 2020.

[52] D. Cooper, D. Apon, Q. Dang, M. Davidson, M. Dworkin, and C. Miller. *Recommendation for Stateful Hash-Based Signature Schemes*. 2019.

[53] D. Coppersmith. "Small Solutions to Polynomial Equations, and Low Exponent RSA Vulnerabilities". In: *Journal of Cryptology* 4 (1997).

[54] J.-S. Coron. "On the Exact Security of Full Domain Hash". In: *CRYPTO*. 2000.

[55] N. Courtois, M. Finiasz, and N. Sendrier. "How to Achieve a McEliece-Based Digital Signature Scheme". In: *ASIACRYPT*. 2001.

[56] N. Courtois, A. Klimov, J. Patarin, and A. Shamir. "Efficient Algorithms for Solving Overdefined Systems of Multivariate Polynomial Equations". In: *EUROCRYPT*. 2000.

[57] J.-M. Couveignes. *Hard Homogeneous Spaces*. Cryptology ePrint Archive, Report 2006/291. 2006.

[58] D. Dachman-Soled, L. Ducas, H. Gong, and M. Rossi. "LWE with Side Information: Attacks and Concrete Security Estimation". In: *CRYPTO*. 2020.

[59] I. Damgård. "On the Randomness of Legendre and Jacobi Sequences". In: *CRYPTO'88*. 1990.

[60] J.-P. D'Anvers, Q. Guo, T. Johansson, A. Nilsson, F. Vercauteren, and I. Verbauwhede. "Decryption Failure Attacks on IND-CCA Secure Lattice-Based Schemes". In: *PKC*. 2019.

[61] J.-P. D'Anvers, A. Karmakar, S. S. Roy, and F. Vercauteren. *SABER*. Tech. rep. National Institute of Standards and Technology, 2019.

[62] J.-P. D'Anvers, M. Rossi, and F. Virdia. "(One) Failure Is Not an Option: Bootstrapping the Search for Failures in Lattice-Based Encryption Schemes". In: *EUROCRYPT*. 2020.

[63] J. D'Anvers, M. Tiepelt, F. Vercauteren, and I. Verbauwhede. "Timing Attacks on Error Correcting Codes in Post-Quantum Schemes". In: *TIS@CCS*. 2019.

[64] J.-P. D'Anvers, F. Vercauteren, and I. Verbauwhede. "The Impact of Error Dependencies on Ring/Mod-LWE/LWR Based Schemes". In: *Post-Quantum Cryptography - 10th International Conference, PQCrypto*. 2019.

[65] L. De Feo and S. D. Galbraith. "SeaSign: Compact Isogeny Signatures from Class Group Actions". In: *EUROCRYPT*. 2019.

[66] L. De Feo, J. Kieffer, and B. Smith. "Towards Practical Key Exchange from Ordinary Isogeny Graphs". In: *ASIACRYPT*. 2018.

[67] C. de Saint Guilhem, L. De Meyer, E. Orsini, and N. P. Smart. "BBQ: Using AES in Picnic Signatures". In: *SAC*. 2019.

[68] T. Debris-Alazard, N. Sendrier, and J.-P. Tillich. "Wave: A New Family of Trapdoor One-Way Preimage Sampleable Functions Based on Codes". In: *ASIACRYPT*. 2019.

[69] T. Debris-Alazard and J.-P. Tillich. "Two Attacks on Rank Metric Code-Based Schemes: RankSign and an IBE Scheme". In: *ASIACRYPT*. 2018.

[70] J. P. Degabriele, A. Lehmann, K. G. Paterson, N. P. Smart, and M. Strefler. "On the Joint Security of Encryption and Signature in EMV". In: *CT-RSA*. 2012.

[71] C. Diem. "The XL-Algorithm and a Conjecture from Commutative Algebra". In: *ASIACRYPT*. 2004.

[72] W. Diffie and M. E. Hellman. "New Directions in Cryptography". In: *IEEE Transactions on Information Theory* 6 (1976).

[73] J. Ding, M.-S. Chen, A. Petzoldt, D. Schmidt, and B.-Y. Yang. *Rainbow*. Tech. rep. National Institute of Standards and Technology, 2019.

[74] J. Ding, X. Xie, and X. Lin. *A Simple Provably Secure Key Exchange Scheme Based on the Learning with Errors Problem*. Cryptology ePrint Archive, Report 2012/688. 2012.

[75] I. Dinur, D. Kales, A. Promitzer, S. Ramacher, and C. Rechberger. "Linear Equivalence of Block Ciphers with Partial Non-Linear Layers: Application to LowMC". In: *EUROCRYPT*. 2019.

[76] L. Ducas, V. Lyubashevsky, and T. Prest. "Efficient Identity-Based Encryption over NTRU Lattices". In: *ASIACRYPT*. 2014.

[77] L. Ducas and P. Q. Nguyen. "Learning a Zonotope and More: Cryptanalysis of NTRUSign Countermeasures". In: *ASIACRYPT*. 2012.

[78] T. ElGamal. "A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms". In: *IEEE Transactions on Information Theory* (1985).

[79] T. Espitau, P.-A. Fouque, B. Gérard, and M. Tibouchi. "Side-Channel Attacks on BLISS Lattice-Based Signatures: Exploiting Branch Tracing against

strongSwan and Electromagnetic Emanations in Microcontrollers". In: *ACM CCS*. 2017.

[80] fail0verflow. "Console Hacking 2010: PS3 Epic Fail". In: *27th Chaos Communications Congress*. 2010.

[81] J. C. Faugère. "A New Efficient Algorithm for Computing GröBner Bases without Reduction to Zero (F5)". In: *ISSAC*. Lille, France, 2002. ISBN: 1581134843.

[82] J. Faugère, V. Gauthier-Umaña, A. Otmani, L. Perret, and J. Tillich. "A Distinguisher for High-Rate McEliece Cryptosystems". In: *IEEE Trans. Inf. Theory* 10 (2013).

[83] L. D. Feo. *Mathematics of Isogeny Based Cryptography*. 2017. arXiv: 1711.04062 [cs.CR].

[84] L. D. Feo, D. Jao, and J. Plût. "Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies". In: *Journal of Mathematical Cryptology* 3 (2014).

[85] A. Fiat and A. Shamir. "How to Prove Yourself: Practical Solutions to Identification and Signature Problems". In: *CRYPTO'86*. 1987.

[86] S. Fluhrer. *Cryptanalysis of ring-LWE based key exchange with key share reuse*. Cryptology ePrint Archive, Report 2016/085. 2016.

[87] G. Fumaroli, A. Martinelli, E. Prouff, and M. Rivain. "Affine Masking against Higher-Order Side Channel Analysis". In: *SAC*. 2011.

[88] P. Gaborit, O. Ruatta, J. Schrek, and G. Zémor. "RankSign: An Efficient Signature Algorithm Based on the Rank Metric". In: *Post-Quantum Cryptography - 6th International Workshop, PQCrypto*. 2014.

[89] S. Galbraith, L. Panny, B. Smith, and F. Vercauteren. *Quantum Equivalence of the DLP and CDHP for Group Actions*. Cryptology ePrint Archive, Report 2018/1199. 2018.

[90] S. D. Galbraith, C. Petit, B. Shani, and Y. B. Ti. "On the Security of Supersingular Isogeny Cryptosystems". In: *ASIACRYPT*. 2016.

[91] T. Gellersen, O. Seker, and T. Eisenbarth. *Differential Power Analysis of the Picnic Signature Scheme*. Cryptology ePrint Archive, Report 2020/267. 2020.

[92] C. Gentry, J. Jonsson, J. Stern, and M. Szydlo. "Cryptanalysis of the NTRU Signature Scheme (NSS) from Eurocrypt 2001". In: *ASIACRYPT*. 2001.

[93] C. Gentry, C. Peikert, and V. Vaikuntanathan. "Trapdoors for hard lattices and new cryptographic constructions". In: *40th ACM STOC*. 2008.

[94] O. Goldreich. "Two Remarks Concerning the Goldwasser-Micali-Rivest Signature Scheme". In: *CRYPTO'86*. 1987.

[95] O. Goldreich, S. Goldwasser, and S. Halevi. "Public-Key Cryptosystems from Lattice Reduction Problems". In: *CRYPTO'97*. 1997.

[96] L. Grassi, C. Rechberger, D. Rotaru, P. Scholl, and N. P. Smart. "MPC-Friendly Symmetric Key Primitives". In: *ACM CCS*. 2016.

[97] L. K. Grover. "A Fast Quantum Mechanical Algorithm for Database Search". In: *28th ACM STOC*. 1996.

[98] T. Güneysu, V. Lyubashevsky, and T. Pöppelmann. "Practical Lattice-Based Cryptography: A Signature Scheme for Embedded Systems". In: *CHES*. 2012.

[99] Q. Guo, T. Johansson, and A. Nilsson. "A Key-Recovery Timing Attack on Post-quantum Primitives Using the Fujisaki-Okamoto Transformation and Its Application on FrodoKEM". In: *CRYPTO*. 2020.

[100] Q. Guo, T. Johansson, and P. Stankovski. "A Key Recovery Attack on MDPC with CCA Security Using Decoding Errors". In: *ASIACRYPT*. 2016.

[101] Q. Guo, T. Johansson, and J. Yang. "A Novel CCA Attack Using Decryption Errors Against LAC". In: *ASIACRYPT*. 2019.

[102] S. Guo, P. Kamath, A. Rosen, and K. Sotiraki. "Limits on the Efficiency of (Ring) LWE Based Non-interactive Key Exchange". In: *PKC*. 2020.

[103] C. Hall, I. Goldberg, and B. Schneier. "Reaction Attacks against several Public-Key Cryptosystems". In: *ICICS 99*. 1999.

[104] M. Hamburg. *Three Bears*. Tech. rep. National Institute of Standards and Technology, 2019.

[105] J. Hoffstein, N. Howgrave-Graham, J. Pipher, J. H. Silverman, and W. Whyte. "NTRUSIGN: Digital Signatures Using the NTRU Lattice". In: *CT-RSA*. 2003.

[106] J. Hoffstein, J. Pipher, and J. H. Silverman. "NSS: An NTRU Lattice-Based Signature Scheme". In: *EUROCRYPT*. 2001.

[107] J. Hoffstein, J. Pipher, and J. H. Silverman. "NTRU: A Ring-Based Public Key Cryptosystem". In: *ANTS*. 1998. ISBN: 3-540-64657-4.

[108] D. Hofheinz, K. Hövelmanns, and E. Kiltz. "A Modular Analysis of the Fujisaki-Okamoto Transformation". In: *TCC*. 2017.

[109] J. Howe, M. Martinoli, E. Oswald, and F. Regazzoni. "Optimised Lattice-Based Key Encapsulation in Hardware". In: *NIST's Second PQC Standardization Conference* (2019).

[110] J. Howe, T. Prest, and D. Apon. *SoK: How (not) to Design and Implement Post-Quantum Cryptography*. Cryptology ePrint Archive, Report 2021. 2021.

[111] J. Howe, T. Prest, T. Ricosset, and M. Rossi. "Isochronous Gaussian Sampling: From Inception to Implementation". In: *Post-Quantum Cryptography - 11th International Conference, PQCrypto*. 2020.

[112] N. Howgrave-Graham, P. Q. Nguyen, D. Pointcheval, J. Proos, J. H. Silverman, A. Singer, and W. Whyte. "The Impact of Decryption Failures on the Security of NTRU Encryption". In: *CRYPTO*. 2003.

[113] W.-L. Huang, J.-P. Chen, and B.-Y. Yang. "Power Analysis on NTRU Prime". In: *IACR TCHES* 1 (2020).

[114] A. Hülsing. "W-OTS+ - Shorter Signatures for Hash-Based Signature Schemes". In: *AFRICACRYPT 13*. 2013.

[115] A. Hulsing, D. J. Bernstein, C. Dobraunig, M. Eichlseder, S. Fluhrer, S.-L. Gazdag, P. Kampanakis, S. Kolbl, T. Lange, M. M. Lauridsen, F. Mendel, R. Niederhagen, C. Rechberger, J. Rijneveld, P. Schwabe, and J.-P. Aumasson. *SPHINCS+*. Tech. rep. National Institute of Standards and Technology, 2019.

[116] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai. "Zero-knowledge from secure multiparty computation". In: *39th ACM STOC*. 2007.

[117] D. Jao, R. Azarderakhsh, M. Campagna, C. Costello, L. De Feo, B. Hess, A. Jalali, B. Koziel, B. LaMacchia, P. Longa, M. Naehrig, J. Renes, V. Soukharev, D. Urbanik, and G. Pereira. *SIKE*. Tech. rep. National Institute of Standards and Technology, 2019.

[118] D. Jao and L. De Feo. "Towards Quantum-Resistant Cryptosystems from Supersingular Elliptic Curve Isogenies". In: *Post-Quantum Cryptography - 4th International Workshop, PQCrypto*. 2011.

[119] S. Jaques, M. Naehrig, M. Roetteler, and F. Virdia. "Implementing Grover Oracles for Quantum Key Search on AES and LowMC". In: *EUROCRYPT*. 2020.

[120] S. Jaques and J. M. Schanck. "Quantum Cryptanalysis in the RAM Model: Claw-Finding Attacks on SIKE". In: *CRYPTO*. 2019.

[121] A. Joux and V. Vitse. *A crossbred algorithm for solving Boolean polynomial systems.* Cryptology ePrint Archive, Report 2017/372. 2017.

[122] D. Kales and G. Zaverucha. *An Attack on Some Signature Schemes Constructed From Five-Pass Identification Schemes.* Cryptology ePrint Archive, Report 2020/837. 2020.

[123] M. J. Kannwischer, P. Pessl, and R. Primas. "Single-Trace Attacks on Keccak". In: *IACR TCHES* 3 (2020).

[124] J. Katz, V. Kolesnikov, and X. Wang. "Improved Non-Interactive Zero Knowledge with Applications to Post-Quantum Signatures". In: *ACM CCS.* 2018.

[125] G. Kuperberg. "A Subexponential-Time Quantum Algorithm for the Dihedral Hidden Subgroup Problem". In: *SIAM J. Comput.* 1 (2005).

[126] G. Kuperberg. "Another Subexponential-time Quantum Algorithm for the Dihedral Hidden Subgroup Problem". In: *TQC.* 2013.

[127] T. Laarhoven, M. Mosca, and J. van de Pol. "Finding shortest lattice vectors faster using quantum search". In: *Des. Codes Cryptogr.* 2-3 (2015).

[128] N. Lahr, R. Niederhagen, R. Petri, and S. Samardjiska. "Side Channel Information Set Decoding Using Iterative Chunking - Plaintext Recovery from the "Classic McEliece" Hardware Reference Implementation". In: *ASIACRYPT.* 2020.

[129] L. Lamport. *Constructing Digital Signatures from a One-way Function.* Technical Report SRI-CSL-98. SRI International Computer Science Laboratory, 1979.

[130] A. Langlois and D. Stehlé. "Worst-case to average-case reductions for module lattices". In: *Des. Codes Cryptogr.* 3 (2015).

[131] F. Liu, T. Isobe, and W. Meier. *Cryptanalysis of Full LowMC and LowMC-M with Algebraic Techniques.* Cryptology ePrint Archive, Report 2020/1034. 2020.

[132] V. Lyubashevsky. "Fiat-Shamir with Aborts: Applications to Lattice and Factoring-Based Signatures". In: *ASIACRYPT.* 2009.

[133] V. Lyubashevsky, L. Ducas, E. Kiltz, T. Lepoint, P. Schwabe, G. Seiler, and D. Stehlé. *CRYSTALS-DILITHIUM.* Tech. rep. National Institute of Standards and Technology, 2019.

[134] V. Lyubashevsky, C. Peikert, and O. Regev. "On Ideal Lattices and Learning with Errors over Rings". In: *EUROCRYPT.* 2010.

[135] R. J. McEliece. "A Public-Key Cryptosystem Based on Algebraic Coding Theory". In: *JPL DSN Progress Report* (1978).

[136] R. C. Merkle. "A Certified Digital Signature". In: *CRYPTO'89.* 1990.

[137] D. Micciancio and C. Peikert. "Trapdoors for Lattices: Simpler, Tighter, Faster, Smaller". In: *EUROCRYPT.* 2012.

[138] V. Migliore, B. Gérard, M. Tibouchi, and P.-A. Fouque. "Masking Dilithium - Efficient Implementation and Side-Channel Evaluation". In: *ACNS 19.* 2019.

[139] K. Mus, S. Islam, and B. Sunar. "QuantumHammer: A Practical Hybrid Attack on the LUOV Signature Scheme". In: *ACM CCS 20.* 2020.

[140] M. Naehrig, E. Alkim, J. Bos, L. Ducas, K. Easterbrook, B. LaMacchia, P. Longa, I. Mironov, V. Nikolaenko, C. Peikert, A. Raghunathan, and D. Stebila. *FrodoKEM.* Tech. rep. National Institute of Standards and Technology, 2019.

[141] P. Q. Nguyen and O. Regev. "Learning a Parallelepiped: Cryptanalysis of GGH and NTRU Signatures". In: *EUROCRYPT.* 2006.

[142] NIST. *Submission Requirements and Evaluation Criteria for the Post-Quantum Cryptography Standardization Process.* 2016.

[143] E. Oswald and S. Mangard. "Template Attacks on Masking - Resistance Is Futile". In: *CT-RSA.* 2007.

[144] A. Park, K.-A. Shim, N. Koo, and D.-G. Han. "Side-Channel Attacks on Post-Quantum Signature Schemes based on Multivariate Quadratic Equations". In: *IACR TCHES* 3 (2018).

[145] K. G. Paterson and R. Villanueva-Polanco. "Cold Boot Attacks on NTRU". In: *INDOCRYPT*. 2017.

[146] C. Peikert. *A Decade of Lattice Cryptography*. Cryptology ePrint Archive, Report 2015/939. 2015.

[147] C. Peikert. "He Gives C-Sieves on the CSIDH". In: *EUROCRYPT*. 2020.

[148] C. Peikert. "How (Not) to Instantiate Ring-LWE". In: *SCN 16*. 2016.

[149] C. Peikert. "Lattice Cryptography for the Internet". In: *Post-Quantum Cryptography - 6th International Workshop, PQCrypto*. 2014.

[150] C. Peikert and Z. Pepin. "Algebraically Structured LWE, Revisited". In: *TCC*. 2019.

[151] A. Perrig. "The BiBa One-Time Signature and Broadcast Authentication Protocol". In: *ACM CCS*. 2001.

[152] P. Pessl, L. G. Bruinderink, and Y. Yarom. "To BLISS-B or not to be: Attacking strongSwan's Implementation of Post-Quantum Signatures". In: *ACM CCS*. 2017.

[153] D. Pointcheval and J. Stern. "Security Proofs for Signature Schemes". In: *EUROCRYPT'96*. 1996.

[154] R. L. V. Polanco. "Cold Boot Attacks on Post-Quantum Schemes". PhD thesis. Royal Holloway, University of London, 2018.

[155] E. Prange. "The use of information sets in decoding cyclic codes". In: *IRE Trans. Inf. Theory* 5 (1962).

[156] T. Prest, P.-A. Fouque, J. Hoffstein, P. Kirchner, V. Lyubashevsky, T. Pornin, T. Ricosset, G. Seiler, W. Whyte, and Z. Zhang. *FALCON*. Tech. rep. National Institute of Standards and Technology, 2019.

[157] P. Ravi, M. P. Jhanwar, J. Howe, A. Chattopadhyay, and S. Bhasin. "Exploiting determinism in lattice-based signatures: practical fault attacks on pqm4 implementations of NIST candidates". In: *AsiaCCS*. 2019.

[158] O. Regev. "On lattices, learning with errors, random linear codes, and cryptography". In: *37th ACM STOC*. 2005.

[159] L. Reyzin and N. Reyzin. "Better than BiBa: Short one-time signatures with fast signing and verifying". In: *Information Security and Privacy 2002*. 2002.

[160] R. L. Rivest, A. Shamir, and L. M. Adleman. "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems". In: *Communications of the Association for Computing Machinery* 2 (1978).

[161] M. Rossi, M. Hamburg, M. Hutter, and M. E. Marson. "A Side-Channel Assisted Cryptanalytic Attack Against QcBits". In: *CHES*. 2017.

[162] A. Rostovtsev and A. Stolbunov. *Public-Key Cryptosystem Based On Isogenies*. Cryptology ePrint Archive, Report 2006/145. 2006.

[163] K. Sakumoto, T. Shirai, and H. Hiwatari. "On Provable Security of UOV and HFE Signature Schemes against Chosen-Message Attack". In: *Post-Quantum Cryptography - 4th International Workshop, PQCrypto*. 2011.

[164] K. Sakumoto, T. Shirai, and H. Hiwatari. "Public-Key Identification Schemes Based on Multivariate Quadratic Polynomials". In: *CRYPTO*. 2011.

[165] S. Samardjiska, M.-S. Chen, A. Hulsing, J. Rijneveld, and P. Schwabe. *MQDSS*. Tech. rep. National Institute of Standards and Technology, 2019.

[166] S. Samardjiska, P. Santini, E. Persichetti, and G. Banegas. "A Reaction Attack Against Cryptosystems Based on LRPC Codes". In: *LATINCRYPT*. 2019.

[167] N. Samwel, L. Batina, G. Bertoni, J. Daemen, and R. Susella. "Breaking Ed25519 in WolfSSL". In: *CT-RSA*. 2018.

[168] C.-P. Schnorr. "Efficient Identification and Signatures for Smart Cards". In: *CRYPTO'89*. 1990.

[169] C. Schnorr and M. Euchner. "Lattice basis reduction: Improved practical algorithms and solving subset sum problems". In: *Math. Program.* (1994).

[170] P. Schwabe, R. Avanzi, J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, G. Seiler, and D. Stehlé. *CRYSTALS-KYBER*. Tech. rep. National Institute of Standards and Technology, 2019.

[171] A. Shamir. "An Efficient Identification Scheme Based on Permuted Kernels (Extended Abstract) (Rump Session)". In: *CRYPTO'89*. 1990.

[172] P. W. Shor. "Algorithms for Quantum Computation: Discrete Logarithms and Factoring". In: *35th FOCS*. 1994.

[173] B.-Y. Sim, J. Kwon, K. Y. Choi, J. Cho, A. Park, and D.-G. Han. "Novel Side-Channel Attacks on Quasi-Cyclic Code-Based Cryptography". In: *IACR TCHES* 4 (2019).

[174] J. Stern. "A New Identification Scheme Based on Syndrome Decoding". In: *CRYPTO'93*. 1994.

[175] J. Stern. "A new paradigm for public key identification". In: *IEEE Trans. Inf. Theory* 6 (1996).

[176] M. Szydlo. "Merkle Tree Traversal in Log Space and Time". In: *EUROCRYPT*. 2004.

[177] R. C. Torres and N. Sendrier. "Analysis of Information Set Decoding for a Sub-linear Error Weight". In: *Post-Quantum Cryptography - 7th International Workshop, PQCrypto*. 2016.

[178] P. C. van Oorschot and M. J. Wiener. "Parallel Collision Search with Cryptanalytic Applications". In: *Journal of Cryptology* 1 (1999).

[179] K. Verhulst. "Power Analysis and Masking of Saber". MA thesis. Belgium: KU Leuven, 2019.

[180] P. Véron. "Improved identification schemes based on error-correcting codes". In: *Appl. Algebra Eng. Commun. Comput.* 1 (1996).

[181] B. Yang and J. Chen. "All in the XL Family: Theory and Practice". In: *ICISC*. 2004.

[182] Y. Yarom and K. Falkner. "FLUSH+RELOAD: A High Resolution, Low Noise, L3 Cache Side-Channel Attack". In: *USENIX Security*. 2014.

[183] G. Zaverucha, M. Chase, D. Derler, S. Goldfeder, C. Orlandi, S. Ramacher, C. Rechberger, D. Slamanig, J. Katz, X. Wang, and V. Kolesnikov. *Picnic*. Tech. rep. National Institute of Standards and Technology, 2019.

[184] M. Zhandry. "A note on the quantum collision and set equality problems". In: *Quantum Inf. Comput.* 7&8 (2015).

[185] Z. Zhang, C. Chen, J. Hoffstein, W. Whyte, J. M. Schanck, A. Hulsing, J. Rijneveld, P. Schwabe, and O. Danba. *NTRUEncrypt*. Tech. rep. National Institute of Standards and Technology, 2019.

Quantifying the Influence of Smoldering Particle Size and Radiant Heat Flux on Ignition of Fuel Beds

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: The magnitude of destruction from fire spread processes are hard to imagine during large outdoor fire outbreaks. Even though these disasters are of great public concern and intense media interest, much knowledge is still needed to validate computational models of the complex ignition processes in large outdoor fires. In this context, wind-driven showers of smoldering wood particles of varying sizes were directed at fuel beds exposed to radiant heat flux**.** The influence of smoldering particle size for a fixed applied radiant heat flux level to a fuel bed was studied to observe subsequent ignition behavior. Results of this study are discussed.

Keywords: Large Outdoor Fires; Firebrands; Ignition

1. Introduction

Devastating large outdoor fires have been responsible for destruction of vast amounts of infrastructure and loss of human life. Wildland fires that spread into urban areas, known as wildland-urban interface (WUI) fires, are capable of enormous destruction [**1**]. WUI fires are distinct from wildland fires; WUI fires include the combustion of both vegetative and human-made fuels and occur where large population centers exist whereas wildland fires include the combustion of vegetative fuels and occur in uninhabited areas. The rise of densely populated urban areas has also seen the development of large urban fires. In addition, the rise of informal settlement communities in Southeast Asia and Africa continues to result in large outdoor fires capable of great destruction.

As a wildland fire reaches an urban area, structure-to-structure fire spread processes will occur via the same mechanisms as those in informal settlement fires and urban fires: radiant heat, direct flame contact, and firebrands. Firebrands, or smoldering and/or flaming particles, are in fact the main culprit to destroy structures in large outdoor fires. In the case of WUI fires, the production of firebrands occurs from the combustion dynamics of vegetative and human-made fuel elements, such as homes and other structures. For urban fires and informal settlement fires, firebrands are produced primarily from human-made fuel elements. Firebrand combustion has a series of important aspects: initial generation or formation from the combustion of both vegetative and structural fuel types, transport, deposition, and ignition of fuel sources generally far removed the original fire source [**2**].

In the combustion process of vegetative fuels, pyrolysis of the fuel elements is an important mechanism. During the vegetative combustion process, wind flow around the fuel elements generates and imposes aerodynamic forces. These forces produce moments and stresses along the fuel elements while pyrolysis simultaneously thermally degrades them and reduces their structural integrity. Firebrands are formed when a critical point is reached, and the fuel elements fracture into smaller pieces and are subsequently entrained in the flow [**3**]. Firebrand generation from structure combustion processes is also an important mechanism to generate firebrands in large outdoor fires. These generation processes are an important research area but are not the focus in the present study.

By far the most often studied aspect of firebrand processes is the transport of these combusting particles [**4**]. After the firebrands are generated, they may be initially lofted by the buoyant large outdoor fire plume and transported in the atmospheric boundary layer. Far less studied is the deposition process of firebrands after they are generated and transported. Once the firebrands are generated and transported within the atmospheric boundary layer, firebrands are deposited and will come in contact with other fuel beds. Firebrands may initiate either a smoldering combustion reaction or a flaming combustion reaction. Most of the fundamental ignition studies have been focused on hot metal particles as a surrogate for firebrands [**5**]. Yet, this is an over-simplification of the problem as firebrands are reacting themselves, and these reactions will influence the nature of subsequent ignition processes for various fuel types.

An additional complication is related to the showers of firebrands produced in large outdoor fires, and with the presence of the combustion of vegetative and structural fuels, there is also radiant heat produced that may only enhance ignition processes. These combined ignition processes have been overlooked and are the subject of this investigation. A custom developed experimental apparatus installed inside unique wind facilities in Japan were utilized to undertake the present investigation. In this study, wind-driven showers of smoldering wood particles of varying sizes were directed at fuel beds exposed to radiant heat flux**.** The influence of smoldering particle size for a fixed applied radiant heat flux level to a fuel bed was studied to observe subsequent ignition behavior.

2. Experimental Description

Firebrand showers were generated using an experimental apparatus designed by the authors known as the reduced-scale continuous feed-firebrand generator. The reduced-scale continuous-feed firebrand generator consisted of two parts: the main body where combustion took place and continuous feeding component. A conveyer was used to feed wood samples continuously into the device and was operated at 1.0 cm/s, and wood samples were put on the conveyer belt at 12.5 cm intervals. Two different wood samples were used in order to generate smoldering firebrands, that is particles with different average sizes. First, Japanese Cypress wood samples were used to produce smoldering particles within the mass and projected area of those similar to particles released from structure combustion [**6**]. The overall dimensions of the Japanese Cypress fuel used for firebrand generation before combustion were 28±8 mm (length), 18±6 mm (width), and 3±1.0 mm (thickness) (average ± standard deviation), respectively. These were provided from a supplier and were oven dried and filtered using a 10 mm mesh. The wood feed rate used here was 80 g/min (mass based) or 160 /min (number based), similar to previous non-radiation assisted ignition studies [**7**]. As smoldering particles were desired, the blower was set to provide an average velocity of 400 cm/s.

As a comparison to firebrands of different particle size, the experimental apparatus was also fed with Douglas-fir wood samples. While details of those experiments are provided elsewhere, they are repeated here for completeness [**8**]. Douglas-fir wood samples 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 samples were used in past studies and have been shown to be within the projected area and the mass of firebrands measured from combusting vegetation (conifers). The wood feed rate used was also 80 g/min (mass based) or 160/min (number based). As smoldering particles were desired, the blower was set to provide an average velocity of 500 cm/s.

The fuel beds used for ignition were 300 mm by 300 mm in size and consisted of the Douglas-fir wood pieces. These were installed inside a mock-up corner assembly lined with calcium

Manzello, Samuel; Suzuki, Sayaka. "Quantifying the Influence of Smoldering Particle Size and Radiant Heat Flux on Ignition of Fuel Beds."
Presented at 58 Japanese Combustion Symposium. December 02, 2020 - December 04, 2020.

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. It was observed that firebrands deposited uniformly on the fuel beds when installed inside the mock-up corner assembly. To further simplify the experiments, these wood pieces in the fuel beds were oven dried to remove moisture at 104 °C. Moisture removal was verified by measuring the temporal variation of mass loss of the wood pieces.

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. It was mounted at a height of 440 mm from the fuel bed surface.

The objective of the experimental setup is to investigate the combined effect of radiative heat flux and firebrands on ignition, and therefore the wind speeds were carefully selected. Based on scoping experiments, firebrands are able to ignite fuel beds of wood pieces within 40 min under both wind speeds without radiative heat flux. Firebrand flux landing on fuel bed was measured via video recording (30 frames per second). The characteristics of firebrands, namely the mass and the projected area of firebrands were compared. As shown in **Fig. 1**, the mass and the projected area of firebrands produced under 4 m/s, 6 m/s, and 8 m/s wind are similar as those are within similar mass and projected area class and the average within uncertainties.

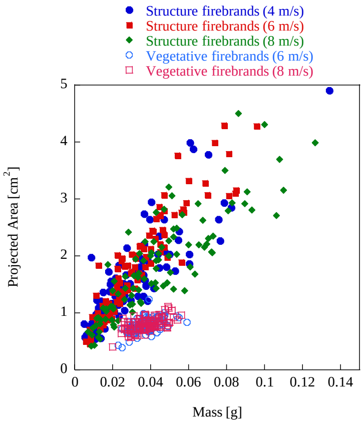


Figure 1 Generated particle sizes based on different wood sample combustion inside the reduced-scale firebrand generator.

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 heat from the radiant panel was 5.76 kW (Quartz-faced Infrared Radiant Panel Heaters, Omega Engineering). The applied radiant heat flux to fuel beds with no wind was 8.5 kW/m^2. In this study, it is important to investigate the coupled effects of firebrand showers and radiant heat flux on fuel bed ignition. Naturally, if the radiant heat flux applied to the fuel beds was too high, firebrands would simply serve as piloted ignition source, as the radiant heat flux would be the dominate ignition mode. Accordingly, a series of experiments using only a radiant heater coupled with spark ignitor was undertaken to determine the limits where radiant heat flux alone in the absence of firebrand showers was capable to ignite the fuel beds used here. At 10 kW/ m^2, the fuel bed was not able to be ignited by a spark within 30 min. For values above 10 kW/m^2, radiant heat flux in the presence of a spark was able to produce ignition of the fuel beds. The radiant heat was measured at multiple points and found to be within 10 %.

The applied wind increases the convective heat loss from the fuel beds, so the measured heat flux to fuel beds, using a total heat flux gauge (Schmidt-Boelter gauge, Hukseflux), under 6 m/s and 8 m/s were 5.6 kW/m^2 and 4.0 kW/m^2, respectively. 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. For pre-heating times of 10 min and 20 min, the wind field was switched on 60 seconds prior to initiating combustion in 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 experiments for the same condition was repeated at least 3 times.

3. Results and Discussion

Fig. 2 displays the mass per unit area for smoldering ignition as function of smoldering particle size, pre-heating time, and applied wind speed. Perhaps the most interesting finding is that as the smoldering particle size is increased, the mass per unit area to initiate smoldering ignition becomes less dependent on the applied wind speed, for a given pre-heat time.

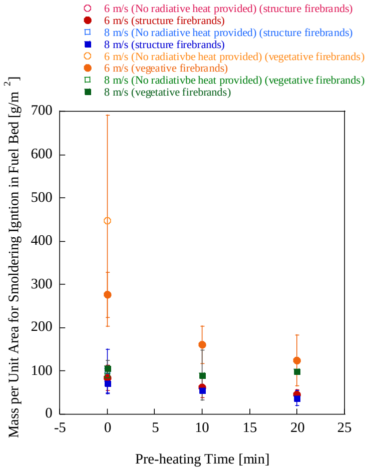


Figure 2 Mass per unit area for smoldering ignition as function of smoldering particle size, pre-heating time, and wind speed.

4. Conclusions

Wind-driven showers of smoldering wood particles of varying sizes were directed at fuel beds exposed to radiant heat flux. As the smoldering particle size is increased, the mass per unit area needed for smoldering ignition in the fuel bed became less dependent on the applied wind speed, for a given pre-heat time.

5. Acknowledgement

This work was partly supported by JSPS KAKENHI (C) Grant Number 18K04675.

References

1. S. L. Manzello *et al*., FORUM Position Paper - The Growing Global Wildland-Urban Interface (WUI) Fire Problem: Priority Needs for Research, *Fire Safety Journal*: 100: 64-66, 2018.
2. S.L. Manzello *et al*., Role of Firebrand Combustion in Large Outdoor Fire Spread, *Prog. Energy Combust. Sci.*, 76: 100801, 2020.
3. B.W. Barr and O. Ezekoye, Thermo-mechanical Modeling of Firebrand Breakage on a Fractal Tree. *Proc. Combust. Inst.*, 34 :2649–2656, 2013.
4. C.S. Tarifa *et al*., On the Flight Paths and Lifetimes of Burning Particles of Wood, *Proc. Combust. Inst.*, 10:1021–1037, 1965.
5. J. Urban, *et al.*, Smoldering Spot Ignition of Natural Fuels by a Hot Metal Particle, *Proc. Combust. Inst.*, 36: 3211- 3218, 2017.
6. S.L. Manzello and S. Suzuki, Generating Firebrand Showers Characteristic of Burning Structures, *Proc. Combust. Inst.*, 36: 3247-3252, 2017.
7. S. Suzuki and S.L. Manzello, Experiments to Provide the Scientific Basis for Laboratory Standard Test Methods for Firebrand Exposure, *Fire Safety Journal* 91: 784-790, 2017.
8. S. Suzuki and S. L. Manzello, Investigating Coupled Effect of Radiative Heat Flux and Firebrand Showers on Ignition of Fuel Beds, *Fire Technology*, on-line, 2020.

Manzello, Samuel; Suzuki, Sayaka. "Quantifying the Influence of Smoldering Particle Size and Radiant Heat Flux on Ignition of Fuel Beds."

Investigating Soot Generated from Engineered Building Materials

SUZUKI, Sayaka[1], MANZELLO, Samuel L.[2]

[1] National Research Institute of Fire and Disaster (NRIFD), Japan
[2] National Instiute of Standards and Technology (NIST), USA

Abstract: Large outdoor fires are of concern all over the world. While fire spread processes are noted for their immediate destruction and loss of life, the generation of particulate matter from the combustion of both vegetative and structural fuels in the event of large outdoor fire outbreaks may have inhalable health effects. In this study, samples of oriented strand board (OSB), a common structural fuel present in the built environment, were ignited using a radiant heater coupled to a spark igniter and soot samples were taken to begin to look at the structure of the generated soot agglomerates. Results of this study will be presented.

Keywords: Large Outdoor Fires; Particulate Emissions

1. Introduction

Throughout the world, devastating large outdoor fires have been responsible for destruction of vast amounts of infrastructure and loss of human life. Across many continents, wildland fires that spread into urban areas, known as wildland-urban interface (WUI) fires, are capable of enormous destruction. In 2018, WUI fires in the state of California destroyed more than 18,000 structures and caused scores of fatalities. WUI fires continue to occur throughout the Americas, Australia, Europe, and Asia. It is important to distinguish WUI fires from wildland fires; WUI fires include the combustion of both vegetative and human-made fuels and occur where large population centers exist, whereas wildland fires include the combustion of vegetative fuels and occur in uninhabited areas.

The rise of densely populated urban areas has also seen the development of large urban fires. In Japan, such fires have occurred for hundreds of years. The most recent of these occurred in the winter of 2016 in Niigata, Japan. Similarly, the USA has also experienced several major urban fires, such as the Great Chicago Fire in 1872 and the Baltimore Fire in 1904. In some cases, earthquakes have served to initiate these fires but it is not a necessary condition for these urban fires to develop. In addition, the rise of informal settlement communities in Southeast Asia and Africa continues to result in large outdoor fires capable of great destruction.

There exist important similarities in the flame spread processes in informal settlement fires, urban fires, and WUI fires. As a wildland fire reaches an urban area, structure-to-structure fire spread processes will occur via the same mechanisms as those in informal settlement fires and urban fires: radiant heat, direct flame contact, and firebrands. These processes are shown in **Fig. 1**.

In large outdoor fire events, the release of combustion products is known to cause severe visibility and health concerns. On a global scale, the combustion of vegetative fuels is believed to be the largest contributor of particulate emissions and the second largest contributor of gaseous emissions [**1**]. During the 2018 WUI fires in Northern California, particulate emissions resulted in a near complete shutdown of the city of San Francisco.

To date, the most common techniques to attempt to quantify emissions from wildland fires have focused on the premise of developing requisite emissions factors (EF) [**1**]. These EFs determine the ratio of a particular species emitted based on a known amount of combusted mass. Accordingly, EFs are reported for CO_2, CO, and $PM_{2.5}$. The range of reported EF does not consider the combustion of structural fuels of importance to WUI fires, urban fires, and informal settlement fires. As a result, in actual situations, not only are the vegetation species responsible for the gaseous and particulate emissions but also the combustion process from structures, automobiles, and other human-made materials only add to the multitude of emissions.

Another limitation is that most of the EF for vegetative fuels are based on prescribed burning, or controlled outdoor burns conducted for various fire management purposes. Prescribed burning has the advantage that it is conducted over realistic scales, but the fire exposure conditions do not mimic what is seen in actual large outdoor fires. As the sheer intensity and scale of actual fire events cannot be conducted safely, these prescribed fires are undertaken, for example, under low ambient wind conditions. Naturally, laboratory-scale experiments do not mimic actual large outdoor fire events as well, but the advantage is that laboratory-scales provide opportunities to benchmark and develop new and improved diagnostic methods that may lead to improved fundamental understanding of the physics of these emissions processes.

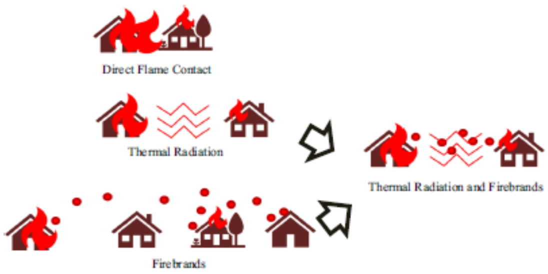


Figure 1 Schematic showing fire spread mechanisms in large outdoor fires.

Better understanding of large outdoor fire gaseous and particulate emissions is an area where fundamental combustion science may play a key role to unravel these processes at the laboratory scale. The combustion community has developed a suite of laser diagnostic capabilities to quantify gaseous and particulate emissions from traditional combustion sources, such as engines and gas turbines [**2-3**]. For example, diagnostic methodologies, such as laser induced incandescence (LII), may be used to quantify particulate emissions from a range of fuel types seen in large outdoor fires. Surprisingly, there is a dearth of measurements on the details of particulate sizes from the combustion processes of both vegetative, structural fuels, and other human-made fuel types. To date, emission measurements from mainly vegetative fuels have been performed using only bulk filter collection techniques, and the full suite of spatially and temporally resolved diagnostics developed by the combustion community remain largely unapplied [**4**].

In this study, samples of oriented strand board (OSB), a common structural fuel present in the built environment, were ignited using a radiant heater coupled to a spark igniter and soot samples were taken to begin to look at the structure of the generated soot agglomerates.

2. Experimental Description

All experiments were conducted at experimental facilities at the National Research Institute of Fire and Disaster (NRIFD). The

Manzello, Samuel; Suzuki, Sayaka. "Investigating Soot Generated from Engineered Building Materials." Presented at 58 Japanese Combustion Symposium. December 02, 2020 - December 04, 2020.

experimental setup consisted of a radiant heater coupled to spark igniter. **Fig. 2** is a schematic of the experimental setup. Samples of OSB were cut into sizes of 100 mm by 100 mm. As commercial samples of OSB was used, the thickness was fixed at 11 mm.

The use of engineered wood products has been common worldwide. In the USA, there has been a dramatic shift to the use of OSB; historically plywood was the dominant material used in roof-sheathing [**5**]. The reasons for this are primarily economic in nature; OSB is manufactured from smaller trees as compared to plywood and consists primarily of wood fragments. Similar trends have been seen in Japan.

As part of this initial proof of concept study, a radiant heat flux of 50 kW/m^2 was applied and the spark was operated continuously. Under these conditions, the OSB samples ignited with sustained flaming ignition within 10 sec.

To sample soot generated, the well-known principle of thermophoretic sampling was used. In the presence of a temperature gradient, the hot soot particles will be collected using cold grids that may be used for Scanning Electron Microscope (SEM) and Transmission Electron Microscopy (TEM) analysis. Here, SEM was used a first step to image the overall structure of the particulate samples. In this study, the sampling time used was varied from 1 sec, to 2 sec, to 3 sec. Namely, the grid was inserted in the flame for these times.

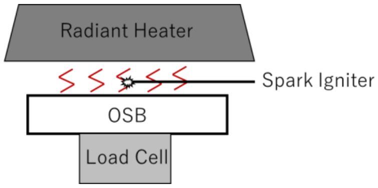


Figure 2 Schematic of the experimental setup.

3. Results and Discussion

Experiments were conducted for one applied heat flux level and all samples were taken at the same time after the onset of sustained flaming combustion of the OSB sample. The total sampling time was varied from 1 sec to 3 sec. **Fig. 3** and **Fig. 4** displays soot agglomerates imaged with the SEM for two different sampling times.

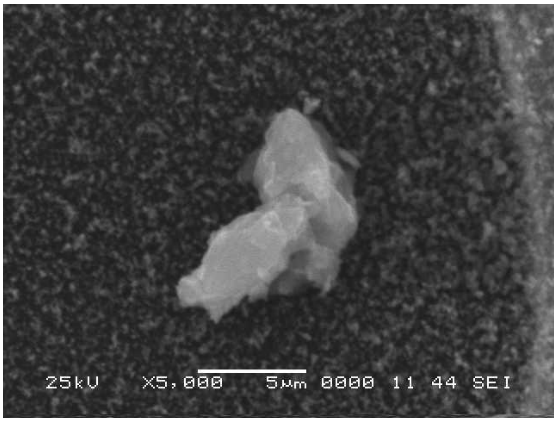


Figure 3 SEM image of soot agglomerate with a grid insertion time of 1 sec.

It is interesting to observe that the structure of the soot agglomerates were not markedly different between the two sampling times.

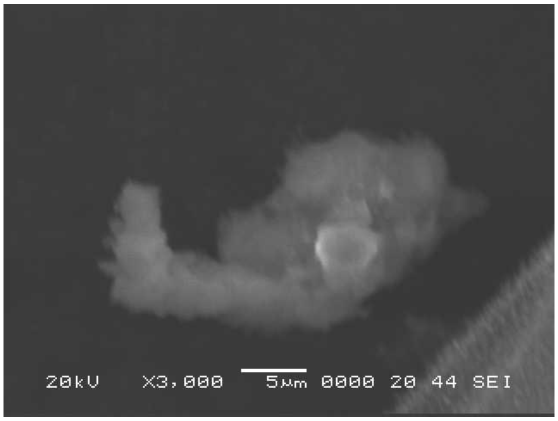


Figure 4 SEM image of soot agglomerate with a grid insertion time of 3 sec.

While the SEM images are useful, TEM analysis is needed to provide finer details of the interior structures of the soot agglomerates and is also better suited to apply image analysis methods to determine important physical characteristics such as primary particle size [**6**]. Obviously, these differences are due to the scanning as opposed to transmission electron microscopy techniques. These analyses are currently in progress using TEM.

4. Conclusions

Fire spread processes are noted for their immediate destruction and loss of life, yet the generation of particulate matter from the combustion of both vegetative and structural fuels in the event of large outdoor fire outbreaks may have inhalable health effects. In this study, samples of oriented strand board (OSB), a common structural fuel present in the built environment, were ignited using a radiant heater coupled to a spark igniter, and soot samples were taken to begin to look at the structure of the generated soot agglomerates using SEM. Although the results are preliminary, structure of soot agglomerates were not markedly different between the two sampling times. Additional analyses are in progress using TEM to better compare the internal structure of the agglomerates and compare these to extensive literature databases on soot formation from various hydrocarbon fuels [**7**].

References

1. S.K. Akagi, *et al.*, Emissions Factors for Open and Domestic Biomass Burning for Use in Atmospheric Models, *Atmos. Chem. Phys*. 11: 4039-4072, 2011.
2. J. Wolfram, Lasers in Combustion: From Basic Theory to Practical Devices, *Proceedings of the Combustion Institute* 27: 1-41, 1998.
3. C. Goldenstein, *et al*., Infrared-Laser Absorption Sensing for Combustion Gases, *Progress in Energy and Combustion Science* 60: 132-176, 2017.
4. Y. Hu, *et al*., Experimental Study on Moisture Content Effects on the Transient Gas and Particle Emissions from Peat Fires, *Combustion and Flame* 209: 408-417, 2019.
5. R.S. White, G. Winandy, Fire Performance of Oriented Strand Board (OBS), Seventeenth Annual BCC Conference on Fire Retardancy, 2006, pp. 297–390.
6. S.L. Manzello and M.Y. Choi, Morphology of Soot Collected in Microgravity Droplet Flames, *International Journal of Heat and Mass Transfer*, 45: 1109-1115, 2002.
7. H. Wang, Formation of Nascent Soot and Other Condensed Phase Materials in Flames, *Proceedings of the Combustion Institute* 33: 41-67, 2011.

Proceedings of the Cold-Formed Steel Research Consortium Colloquium
20-22 October 2020 (cfsrc.org)

Application of ASCE 41 to a two-story CFS building

M.S. Speicher[1], Z. Zhang[2], B.W. Schafer[3]

Abstract

The objective of this paper is to summarize the evaluation results from applying the updated performance-based seismic design provisions, ASCE 41-17, on a cold-formed steel framed building sited in a location with high seismic demands. The assessment included examination of the existing design and consideration of the retrofits required to bring the design into compliance with ASCE 41-17. The assessment of the building relies on the linear procedures and m-factors in ASCE 41-17 and follows the same basic process as the original design. Despite the fact that the studied building is compliant with ASCE 7 and AISI S400, and successfully withstood shake table testing in excess of maximum considered earthquake levels with no permanent damage and no residual drift, ASCE 41-17 finds the building to be deficient. The work highlights that, for cold-formed steel framing, even though ASCE 41 is based on the same tested shear walls that ASCE 7/AISI S400 rely upon, the component-based procedures of ASCE 41 do not easily account for the larger system overstrength and ductility that are included and validated for actual systems. Further work is needed to improve ASCE 41 to account for full system performance, this is particularly important given ASCE 41's growing role as the benchmark performance-based standard for seismic assessment and design.

1. Introduction

ASCE 41-17 [1] provides seismic design procedures for assessment, repair/retrofit, and new building design. Recent studies have highlighted critical differences between ASCE 41-based seismic design and conventional seismic design using ASCE 7 and materials standards (e.g, AISC 341) [2-11]. In the last code update cycle the ASCE 41 provisions were significantly expanded to reflect the best available data for cold-formed steel response [13]. The impact of these changes, and comparisons between ASCE 7-based design and ASCE 41-based design, do not currently exist for cold-formed steel (CFS) framed buildings. To complete such a comparison the two-story CFS framed building designed and tested during a National Science Foundation (NSF) sponsored George E. Brown Network for Earthquake Engineering Simulation (NEES) research project know as CFS-NEES (Figure 1) is selected as a case study.

The CFS-NEES building was designed to contemporary practice using ASCE 7/AISI S400 and subjected to shake table testing at the University at Buffalo in 2013. The building response was excellent, with only minor damage even for seismic excitations in excess of ASCE 7's maximum considered earthquake levels [14]. Subsequent nonlinear time history analyses further demonstrated that while the building was efficiently designed with respect to ASCE7/AISI S400 (i.e. demand/capacity ratios for the shear walls near 1.0) the building had substantial strength reserve and more than acceptable collapse probabilities [15,16].

[1] Research Structural Engineer, National Institute of Standards and Technology: matthew.speicher@nist.gov
[2] Graduate Research Assistant, Dept. of Civil and Systems Engineering, Johns Hopkins University: zhidongzhang@jhu.edu
[3] Professor, Dept. of Civil and Systems Engineering, Johns Hopkins University: schafer@jhu.edu

Speicher, Matthew; Zhang, Zhidong; Schafer, Benjamin. "Application of ASCE 41 to a two-story CFS building." Presented at Cold-Formed Steel Research Consortium Colloquium. October 20, 2020 - October 22, 2020.

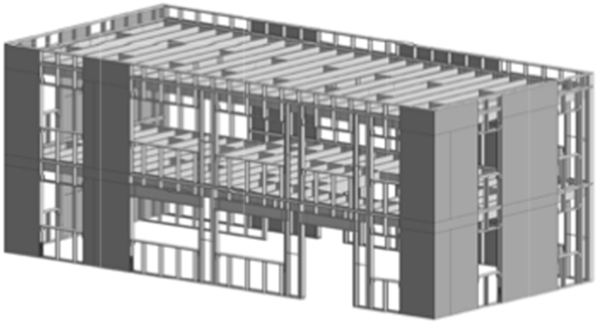


Figure 1. Isometric of framing for 2-story CFS-NEES building (sheathing depicted only on shear walls) [17].

2. Methodology

The original CFS-NEES building design was completed per ASCE 7-05, AISI S100-07, and AISI S213-07 as detailed in [17]. The design was updated to satisfy the latest standards, ASCE 7-16 [18], AISI S100-16 [19], and AISI S400-15 [20]. Then, the updated design was evaluated as an *existing building* using the linear static procedure of ASCE 41-17. Per ASCE 41, the existing building was evaluated for life safety (LS) at the basic safety earthquake (BSE)-1E level and collapse prevention (CP) at the BSE-2E level, where the letter "E" signifies "existing." Only the LS results are provided here, see the complete report in [21] for CP results and full details of the methods, results, and discussion. After the existing building evaluation was completed, a *retrofit* that satisfies the ASCE 41 linear requirements was completed.

2.1 General Approach for ASCE 41 Assessment

ASCE 41 has several different assessment options, from a tier 1 evaluation, which includes a "checklist" cursory style screening, to a tier 3 evaluation, which consists of varying degrees of engineering analysis, with the most complex being the nonlinear dynamic procedure. For this study, the linear static procedure is used, which is the "simplest" form of a tier 3 analysis. The linear static procedure aligns well with the equivalent static force procedure used in traditional design and involves applying an unreduced lateral load, distributed at each story, and then comparing the force demand to the product of the expected capacity and a component capacity modification (*m*)-factor that accounts for the ductility at the selected structural performance level.

2.2 Demand

The first step taken for the ASCE 41 linear assessment is to calculate the demands on the shear walls. The shear walls are considered deformation-controlled components. The base shear of the building that the shear walls must carry, *V*, is calculated from ASCE 41-17 Equation 7-21:

$$V = C_1 C_2 S_a W \tag{1}$$

where C_1 is a modification factor relating expected maximum inelastic displacements to displacements obtained from linear elastic response; C_2 is the modification factor representing the effects of pinched hysteresis shape, cyclic stiffness degradation, and strength deterioration on maximum response; S_a is the response spectrum acceleration at the fundamental period of the building; and W is the effective seismic weight of the building. For the assessment in this study, the approximate value for the product of C_1C_2 is employed from ASCE 41-17 Table 7-3 and is equal to 1.4.

The base shear is distributed to each floor as a static force and the story shears are then distributed to the two sides of the building (1/2 to each side) and then to each shear wall along a side of the building based on their calculated relative stiffness. These individual shear wall demands are used as the deformation-controlled component demands per unit length, v_{ud}, for the assessment.

The demands on the chord studs and ties/hold-downs are determined by treating them as force-controlled components. The "capacity design" approach in ASCE 41 is employed in which the expected capacity of the shear wall is used to determine the maximum forces that can be delivered to the force-controlled components. The required axial load, P_r, and the required moment, M_r, are generated assuming the shear wall is carrying its expected capacity in combination with the appropriate gravity load. The chord studs are subjected to eccentric loads, primarily due to gravity loads framing into the interior flange of the stud from the ledger.

For linear procedures, the combination of actions resulting from dead load (Q_D) and live load (Q_L) with the seismic load (Q_E) follows per ASCE 41-17 Eq. (7-1), adapted here as follows:

$$Q = Q_E + 1.1(Q_D + Q_L) \tag{2}$$

where Q_D is the action resulting from the dead load and Q_L is the action resulting from the live load. Further, Q_L is defined as 25 % of the unreduced live load from ASCE 7. The maximum axial forces in the ties and hold-downs are determined similarly – considering the expected capacity of the shear wall, and considering the case of counteracting loads where ASCE 41-17 Eq. 7-2 holds, adapted here as:

$$Q = Q_E + 0.9(Q_D) \tag{3}$$

2.3 Capacity

The shear wall expected capacity per unit length, v_{ce}, is:

$$v_{ce} = \phi v_n \tag{4}$$

where ϕ is set to 1.0 and v_n is the nominal shear wall capacity per unit length. The nominal shear wall capacity is determined from AISI S400-15. Additionally, the *m*-factors in ASCE 41 can be considered part of the capacity of the shear wall. The *m*-factors are found in ASCE 41-17 Table 9-9 for CFS components. CFS shear walls sheathed with oriented strand board (OSB), considered as primary components, have *m*-factors of 2.5 for life safety (LS) and 3.3 for collapse prevention (CP).

The chord studs are considered force-controlled components, therefore lower-bound strengths are used in the assessment. The lower-bound axial (P_{CL}) and flexural strength (M_{CL}) for the chord studs as specified in ASCE 41-17 Section 9.3.2.3.2 and result in P_{CL}=0.94P_n and M_{CL}=0.94M_n as detailed in [21].

2.4 Acceptance Criteria Check

The linear acceptance criteria check for the shear walls follows the requirements for deformation-controlled components in ASCE 41. With the demand and capacity determined, the linear procedure acceptance criteria for the shear walls is:

$$\frac{v_{ud}}{\kappa v_{ce}} < m \tag{5}$$

where κ is assumed as 1.0 herein. The acceptance criteria check for the chord studs and ties/hold downs follow the requirements for force-controlled components in ASCE 41. The acceptance criteria for the chord studs can be written as the following interaction equation:

$$\kappa\left(\frac{P_{UF}}{P_{CL}} + \frac{M_{UF}}{M_{CL}}\right) \leq 1.0 \quad (6)$$

where P_{CL} is the lower-bound capacity of the chord stud in compression, M_{CL} is the lower-bound capacity of the chord stud in flexure, P_{UF} is the maximum axial load that can be developed in the chord stud due to the shear wall reaching its expected capacity (in combination with dead and live load), and M_{UF} is the flexural load resulting from eccentricity in the loads being delivered to the chord stud. Note, M_{UF} includes second order effects and may be approximated as B_1M_{UF1} where B_1 is the approximate moment magnifier (Equation C1.2.1.1-3 in AISI S100-16 [19]) and M_{UF1} is the first-order demand. The acceptance criteria check for ties/hold-downs is:

$$\kappa\left(\frac{T_{UF}}{T_{CL}}\right) \leq 1.0 \tag{7}$$

where T_{CL} is the lower-bound tension or compression capacity and T_{UF} is the demand arising from the shear wall reaching its expected capacity.

3. Existing Building Evaluation per ASCE 41-17

Per ASCE 41-17 the CFS-NEES building's linear static procedure assessment results for the life safety (LS) performance level at the BSE-1E earthquake hazard level are shown in Table 1.

Shear walls with $v_{ud}/v_{ce} > m$ fail the assessment and are designated with bold and underline. For the 2nd story, 6 out of 10 shear walls fail the assessment. For the first story, 9 out of 10 shear walls fail the assessment.

4. Retrofit Design Evaluation per ASCE 41-17

Each shear wall was individually retrofitted to pass the ASCE 41 assessment. The easiest retrofit option was to increase the number of fasteners. The original fastener spacing was 6 in. (150 mm), therefore for practical purposes a 3 in. (75 mm) fastener spacing was first investigated. If a 3 in. (75 mm) spacing did not give the necessary capacity, double sheathing (i.e. sheathing on both sides of the wall) was the next option examined. If with double-sided sheathing the capacity was sufficient to relax back from 3 in. fastener spacing to 6 in. (150 mm) fastener spacing, then this was done. After iterating through the different options, each shear wall was retrofitted and Table 2 summarizes the results for life safety (LS) at the BSE-1E level.

Required changes for the 1st story shear wall retrofit are significant – the South and East wall lines require double-sided sheathing as does the longest shear wall on the West facing wall line, L1W3. All 1st story shear walls need additional fasteners placed between all existing fasteners to decrease the fastener spacing down to 3 in. (75 mm) The 2nd story shear walls require double-sided sheathing in the same locations as the 1st story, but the existing 6 in. (150 mm) fastener spacing is adequate. The required retrofits would be costly; however, they do not require an increase in shear wall length, thus practically they could be accomplished.

Given the increased capacity of the shear walls due to decreasing the fastener spacing and/or adding sheathing, the chords studs need to be evaluated to determine if they have sufficient capacity to carry the forces created when the shear walls are loaded to their new expected capacity. At the life safety BSE-1E hazard level the existing 2nd story chord studs are adequate for the retrofit, but none of the existing 1st story chord studs are adequate. Retrofit options are possible, but none are without complication. Retrofit designs consisting of adding one or two additional studs to the chord studs are provided for the life safety BSE-1E hazard level in Table 3. The results of the interaction equation are also provided in the tables. In the reported retrofit designs an interaction expression as high as 1.05 was allowed. At the life safety BSE-1E hazard level adding one additional stud (for a total of 3) is found to be sufficient.

The retrofit design requires increased capacity of the shear walls and this also potentially influences the existing story-to-story ties and the hold-down anchorage. The demands, T_{UF}, consider the load combination for counteracting loads. At the life safety BSE-1E hazard level the existing 1st-to-2nd story ties are adequate, but none of the foundation-to-1st story hold-downs are adequate.

Retrofit of the foundation-to-1st story hold-down can be completed relatively simply if a second hold-down (added to the opposite face of the stud) is adequate for the demand. It is possible to place hold-downs side by side as well, thus having as many as 4 commercial hold-downs connected to a built-up chord stud. Non-commercial options using heavy angles are also possible for higher demands. As higher capacity hold-downs are employed, one must note that the anchor bolt sizes typically increase, requiring additional re-design for the retrofit. Capacity of the underlying foundation, particularly with multiple anchors in close proximity, may further limit the available tensile capacity and require additional, more costly and more complex, retrofit. Where possible it is recommended to simply double up the existing S/HDU 6 hold-downs. However, this is not adequate for all the hold-downs in the South walls and East walls and in the L1W3 West wall. For these cases 2 x S/HDU9 hold-downs are specified. These hold-downs have 64 % more strength than the S/HDU6 when connected to 54 mil studs, but require a 7/8 in. (22 mm) anchor bolt.

Table 1. Linear static assessment results of the shear walls considering life safety (LS) at the BSE-1E earthquake hazard level, where plf is pounds per linear foot.

2nd Story				m-factor
Shear wall[a]	v_{ud} plf (kN/m)	v_{ce} plf (kN/m)	v_{ud} / v_{ce}	
L2S1	2039(29.76)	622(9.08)	**3.28**	2.5
L2S2	2623(38.28)	700(10.2)	**3.75**	2.5
L2S3	2039(29.76)	622(9.08)	**3.28**	2.5
L2N1	1684(24.58)	700(10.2)	2.41	2.5
L2N2	1152(16.81)	700(10.2)	1.65	2.5
L2W1	1408(20.55)	622(9.08)	2.26	2.5
L2W2	1408(20.55)	622(9.08)	2.26	2.5
L2W3	2595(37.87)	700(10.2)	**3.71**	2.5
L2E1	1755(25.61)	700(10.2)	**2.51**	2.5
L2E2	2362(34.47)	700(10.2)	**3.37**	2.5

1st Story				m-factor
Shear wall[a]	v_{ud} plf (kN/m)	v_{ce} plf (kN/m)	v_{ud} / v_{ce}	
L1S1	3465(50.57)	733(10.7)	**4.73**	2.5
L1S2	4434(64.71)	825(12.0)	**5.37**	2.5
L1S3	3465(50.57)	733(10.7)	**4.73**	2.5
L1N1	2860(41.74)	825(12.0)	**3.47**	2.5
L1N2	1946(28.40)	825(12.0)	2.36	2.5
L1W1	2401(35.04)	733(10.7)	**3.27**	2.5
L1W2	2401(35.04)	733(10.7)	**3.27**	2.5
L1W3	4383(63.97)	825(12.0)	**5.31**	2.5
L1E1	2979(43.48)	825(12.0)	**3.61**	2.5
L1E2	4002(58.40)	825(12.0)	**4.85**	2.5

Note: **bold and underline** component fails assessment. **a**. shear walls are identified by level one (L1) or two (L2) by face of the building north (N), south (S), east (E), and west (W) - the south (long) and east (short) walls are shown in Figure 1, and finally shear wall number 1, 2, or 3.

Table 2. CFS-NEES retrofit for shear walls for life safety (LS) at the BSE-1E earthquake hazard level, where *s* is the fastener spacing.

	Retrofit				
SW	OSB Sheathing in (mm)	sides	*s* in (mm)	v_{ud} / v_{ce}	m-factor
2nd Story					
L2S1	7/16(11)	**2**	6(150)	1.64	2.5
L2S2	7/16(11)	**2**	6(150)	1.86	2.5
L2S3	7/16(11)	**2**	6(150)	1.64	2.5
L2N1	7/16(11)	1	6(150)	2.41	2.5
L2N2	7/16(11)	1	6(150)	1.65	2.5
L2W1	7/16(11)	1	6(150)	1.85	2.5
L2W2	7/16(11)	1	6(150)	1.85	2.5
L2W3	7/16(11)	**2**	6(150)	2.06	2.5
L2E1	7/16(11)	**2**	6(150)	1.26	2.5
L2E2	7/16(11)	**2**	6(150)	1.69	2.5
1st Story					
L1S1	7/16(11)	**2**	**3(75)**	1.26	2.5
L1S2	7/16(11)	**2**	**3(75)**	1.43	2.5
L1S3	7/16(11)	**2**	**3(75)**	1.26	2.5
L1N1	7/16(11)	1	**3(75)**	1.85	2.5
L1N2	7/16(11)	1	**3(75)**	1.26	2.5
L1W1	7/16(11)	1	**3(75)**	1.73	2.5
L1W2	7/16(11)	1	**3(75)**	1.73	2.5
L1W3	7/16(11)	**2**	**3(75)**	1.43	2.5
L1E1	7/16(11)	**2**	**3(75)**	0.96	2.5
L1E2	7/16(11)	**2**	**3(75)**	1.30	2.5

Note: **bold** indicates changes from original design (7/16" (11 mm) OSB on 1 side with fasteners spaced at 6 in. (152 mm) o.c.).

Table 3. Linear static procedure assessment results of the chord studs considering expected capacities from shear walls retrofitted to meet life safety (LS) at the BSE-1E earthquake hazard level.

1st Story			
SW	Existing Chord Stud	Retrofit Chord Stud	Int'n
L1S1	(2) 600S162-54	**(3) 600S162-54**	1.00
L1S2	(2) 600S162-54	**(3) 600S162-54**	1.03
L1S3	(2) 600S162-54	**(3) 600S162-54**	0.99
L1N1	(2) 600S162-54	**(3) 600S162-54**	0.96
L1N2	(2) 600S162-54	**(3) 600S162-54**	0.94
L1W1	(2) 600S162-54	**(3) 600S162-54**	0.82
L1W2	(2) 600S162-54	**(3) 600S162-54**	0.82
L1W3	(2) 600S162-54	**(3) 600S162-54**	1.01
L1E1	(2) 600S162-54	**(3) 600S162-54**	1.01
L1E2	(2) 600S162-54	**(3) 600S162-54**	1.01

Note: **bold** indicates changes from original design. Interaction allowed up to 1.05 by engineering judgment

6. Discussion

For the studied CFS-framed building, ASCE 41-17 provides a more pessimistic estimation of the seismic response than ASCE 7-16. ASCE 41's *m*-factors are based on direct shear wall tests (as described in [13]) and are ostensibly a more direct and rational gauge of expected behavior than the R and Ω_o factors of ASCE 7, which are based more on experience and judgment than on direct testing [23]. However, in the case of the studied CFS-NEES building, direct testing of the entire building system was conducted and indicated behavior far better than ASCE 7's prediction – even at excitations in excess of the ASCE 7 maximum considered earthquake (MCE)-level, minimal damage occurred [14]. Thus, the true behavior is better than ASCE 7-16's prediction, and far better than ASCE 41-17's prediction.

Subsequent analysis indicated that repetitively framed buildings, such as the CFS-NEES building, have significant overstrength, even more than the amount attributed at Ω_o levels [15]. Examination of the ASCE 7 seismic response modification factors using the FEMA P695 [23] procedure for the CFS-NEES building indicated that if only the shear walls were considered (as essentially is done in ASCE 41 if gravity and non-structural wall contributions to lateral capacity are ignored), then the collapse probabilities are unacceptable. In contrast, if the shear walls and all the gravity framing (unsheathed) were considered, then the collapse probabilities were acceptable – suggesting ASCE 7 response modification factors (R and Ω_o) are justified. Moreover, if the final building, with sheathing, non-structural walls, and finish systems, was considered, then the collapse probabilities were acceptable by an even wider margin and the structural analysis was in line with the shake table test results [16]. Essentially, for this building, and likely this building system type, ASCE 41's lack of an "easy switch" to account for system overstrength in the linear assessment procedure is an important reason that it's linear analysis method provides such pessimistic predictions of performance.

The use of nonlinear static or nonlinear dynamic procedures could provide further insight on the predicted behavior of the building. However, the use of nonlinear procedures is not expected to change the fundamental findings herein: ASCE 41 predicts higher demands than ASCE 7, especially for short period buildings, and does not readily provide a means to easily include system overstrength, thus resulting in conservative assessment outcomes. One proviso on this conclusion, if the gravity and non-structural wall elements are modeled as being meaningfully capable of resisting lateral demands and a rational approach can be adopted for their strength and stiffness degradation, then it is possible, within the ASCE 41 framework, to include the system overstrength. However, where ASCE 7 allows the engineer to include this overstrength effect through a single Ω_o factor, ASCE 41 would require explicit modeling, with significant uncertainty in the parameters, to include the same phenomena.

7. Conclusions

A two-story cold-formed steel framed building, previously designed to ASCE 7 and successfully tested on shake tables in the laboratory, was examined to determine necessary changes if ASCE 41 is adopted for assessment. The two-story cold-formed steel framed building, designed to satisfy ASCE 7, fails when assessed as an existing building per ASCE 41. Retrofit of the two-story cold-formed steel framed building such that it meets the criteria of ASCE 41 essentially requires doubling the capacity of the seismic force resisting system beyond that of ASCE 7. This doubling in capacity is not justified by the experimentally and numerically validated performance of the building. Two primary factors contribute to the conservative nature of ASCE 41's predictions: (1) the basic seismic demands are significantly greater in ASCE 41 than in ASCE 7, especially for short period structures, and (2) large system overstrength, common in repetitively-framed structures, is accounted for in ASCE 7, but not easily in the linear procedures of ASCE 41. Though overstrength may be addressed in ASCE 41 by the higher tier analysis methods (i.e., nonlinear methods), for normal low-rise CFS buildings, this level of effort may not be a realistic option. For ASCE 41 to realize its performance-based design vision and for society to benefit from the flexibility afforded by such frameworks, the basic predicted seismic response for cold-formed steel framed buildings needs to be more closely aligned with reality as demonstrated by shake table tests. Thus, improvements in both demand and capacity procedures for ASCE 41 are needed for this class of building.

8. Acknowledgments and Disclaimers

The contributions of Ivana Olivares are gratefully acknowledged. 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] ASCE (2017) Seismic Evaluation and Retrofit of Existing Buildings (American Society of Civil Engineers).

[2] 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, Gaithersburg, MD), NIST Technical Note 1863-1. Available at http://dx.doi.org/10.6028/NIST.TN.1863-1

[3] Harris JL, Speicher MS (2015) Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings Volume 2: Special Concentrically Braced Frames. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Technical Note 1863-2. Available at http://dx.doi.org/10.6028/NIST.TN.1863-2

[4] Harris JL, Speicher MS (2015) Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings Volume 3: Eccentrically Braced Frames. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Technical Note 1868-3. Available at http://dx.doi.org/10.6028/NIST.TN.1863-3

[5] Speicher MS, Harris JL (2016) Collapse Prevention Seismic Performance Assessment of New Eccentrically Braced Frames using ASCE 41. Engineering Structures 117:344–357. https://doi.org/10.1016/j.engstruct.2016.02.018

[9] Speicher MS, Harris JL (2016) Collapse prevention seismic performance assessment of new special concentrically braced frames using ASCE 41. Engineering Structures 126:652–666. https://doi.org/10.1016/j.engstruct.2016.07.064

[10] Speicher MS, Harris JL (2018) Collapse Prevention seismic performance assessment of new buckling-restrained braced frames using ASCE 41. Engineering Structures. 164:274–289. https://doi.org/10.1016/j.engstruct.2018.01.067

[11] John Harris, Matthew S. Speicher (2018) Assessment of Performance-Based Seismic Design Methods in ASCE 41 for New Steel Buildings: Special Moment Frames. Earthquake Spectra 34(3):977–999. https://doi.org/10.1193/050117EQS079EP

[13] Ayhan D, Rob L. Madsen, Ben W. Schafer (2016) Progress in the Development of ASCE 41 for Cold-Formed Steel. Proceedings of the 23rd International Specialty Conference on Cold-Formed Steel Structures (Baltimore, MD), pp 417–432.

[14] Peterman Kara D., Stehman Matthew J. J., Madsen Rob L., Buonopane Stephen G., Nakata Narutoshi, Schafer Benjamin W. (2016) Experimental Seismic Response of a Full-Scale Cold-Formed Steel-Framed Building. I: System-Level Response. Journal of Structural Engineering 142(12):04016127. https://doi.org/10.1061/(ASCE)ST.1943-541X.0001577

[15] Leng J, Peterman KD, Bian G, Buonopane SG, Schafer BW (2017) Modeling seismic response of a full-scale cold-formed steel-framed building. Engineering Structures 153:146–165. https://doi.org/10.1016/j.engstruct.2017.10.008

[16] Leng J, Buonopane SG, Schafer BW (2020) Incremental dynamic analysis and FEMA P695 seismic performance evaluation of a cold-formed steel–framed building with gravity framing and architectural sheathing. Earthquake Engineering & Structural Dynamics 49(4):394–412. https://doi.org/10.1002/eqe.3245

[17] Rob L. Madsen, N. Nakata, Ben W. Schafer (2013) CFS-NEES Building Structural Design Narrative. Available at http://jhir.library.jhu.edu/handle/1774.2/40584

[18] ASCE (2017) Minimum Design Loads for Buildings and Other Structures (American Society of Civil Engineers).

[19] AISI S100 (2016). Specification for the Design of Cold-Formed Steel Structural Members. American Iron and Steel Institute. Washington, DC.

[20] AISI S400 (2015). North American Standard for Seismic Design of Cold-Formed Steel Structural Systems. American Iron and Steel Institute. Washington, DC.

[21] Speicher, M., Olivares, I., Schafer, B.W. (2020). Seismic Evaluation of a 2-Story Cold-Formed Steel Framed Building using ASCE 41-17. NIST Technical Note 2116, Gaithersburg, MD https://doi.org/10.6028/NIST.TN.2116

[22] FEMA (2009) Quantification of Building Seismic Performance Factors (Department of Homeland Security).

SiF 2020– The 11th International Conference on Structures in Fire

The University of Queensland, Brisbane, Australia, November 30-December 2, 2020 (Author's Copy)

EXPERIMENTAL STUDY ON FIRE RESISTANCE OF A FULL-SCALE COMPOSITE FLOOR ASSEMBLY IN A TWO-STORY STEEL FRAMED BUILDING

Lisa Choe[1,*], Selvarajah Ramesh[2], Xu Dai[3], Matthew Hoehler[4], Matthew Bundy[5]

ABSTRACT

This paper presents the first of four planned fire experiments of a full-scale two-story steel framed building constructed at the National Fire Research Laboratory. This first experiment was aimed to quantify the fire resistance and behaviour of the steel composite floor system commonly built in the United States, incorporating prescriptive approaches for a 2-hour fire resistance rating. The 9.1 m × 6.1 m composite floor assembly, situated in the edge bay on the first floor of the prototype building, was tested to failure under a natural gas fuelled compartment fire and simultaneously applied mechanical loads. The test showed that the protected floor beams and girders of the test assembly achieved matching or superior fire resistance based on the acceptance criteria of standard furnace testing. However, the test floor slab exhibited a potential fire hazard within a specified fire rating period because of the use of a minimum code-compliant shrinkage reinforcement (59 mm^2/m). The heated slab cracked around the interior edges of the test bay less than 30 min into heating, followed by centre cracks along the secondary beam at 70 min. This centre breach, accompanied by ruptures of wire reinforcement, was caused by tension (due to catenary action) developed along the shorter span of the test floor assembly. This result suggests that the minimum slab reinforcement prescribed for normal conditions may not be sufficient to activate tensile membrane action of a composite floor system under the 2-hour standard fire exposure.

Keywords: Composite floors; steel buildings; fire resistance; compartment fire experiments

1 INTRODUCTION

There is lack of experimental data quantifying the fire performance of full-scale composite steel frames designed in accordance with United States (U.S.) building codes and specifications. Standard fire testing of the full-scale composite floor assemblies, incorporating both steel member connections and slab continuity, are extremely rare due to the size limitations of testing furnaces. The National Institute of Standards and Technology (NIST) is conducting a multi-year experimental test program to quantify the behaviour and limit states of full-scale structural steel frames with composite floors under compartment fire conditions. This test program consists of two parts: Phase 1 covers 12.8 m long composite beams with

[1] Research Structural Engineer, National Institute of Standards and Technology (NIST)
e-mail: lisa.choe@nist.gov, ORCID: https://orcid.org/0000-0003-1951-2746 (corresponding author)

[2] Foreign Guest Researcher, National Institute of Standards and Technology (NIST)
e-mail: selvarajah.ramesh@nist.gov, ORCID: https://orcid.org/0000-0002-9525-6767

[3] Foreign Guest Researcher, National Institute of Standards and Technology (NIST)
e-mail: xu.dai@nist.gov, ORCID: https://orcid.org/0000-0002-9617-7681

[4] Research Structural Engineer, National Institute of Standards and Technology (NIST)
e-mail: matthew.hoehler@nist.gov, ORCID: https://orcid.org/0000-0002-6049-7560

[5] Supervisory Mechanical Engineer, National Institute of Standards and Technology (NIST)
e-mail: matthew.bundy@nist.gov, ORCID: https://orcid.org/0000-0002-1138-0307

Choe, Lisa; Ramesh, Selvarajah; Dai, Xu; Hoehler, Matthew; Bundy, Matthew. "Experimental Study on Fire Resistance of a Full-Scale Composite Floor Assembly in a Two-Story Steel Framed Building." Presented at 11th International Conference on Structures in Fire (SiF 2020). November 30, 2020 - December 02, 2020.

simple shear connections [1] and Phase 2 covers 9.1 m by 6.1 m composite floor assemblies in a two-story structural steel gravity frame. This paper presents the first experiment of the Phase 2 study conducted at NIST's National Fire Research Laboratory (NFRL) [2] on November 14, 2019. The test floor assembly was located on the first floor south-edge bay of the two-story prototype building and tested to failure under combined mechanical loads and a standard fire exposure simulated using natural gas burners. The objective of this test was to measure the structural and thermal responses of the composite floor assembly designed and constructed following the current U.S. prescriptive approach and to evaluate its system-level fire resistance based on the American Society of Testing and Materials (ASTM) standard E119 criteria [3]. The experimental results presented herein will serve as baseline to compare with the remaining experiments in Phase 2 and be used to guide the validation of computational models and design tools.

2 TEST STRUCTURE

The two-story structural steel frame was constructed to conduct a series of single-bay compartment fire experiments; see Figure 1(a). This test frame consisted of three bays by two bays in plan with a total floor area of 18 m × 11 m. The total height was approximately 7.2 m. The composite floor assembly was constructed on the first floor (3.8 m above the ground floor), whereas on the second floor a beam framing was erected without concrete slab. The same sizes of wide-flange steel shapes were used for both the first and second floor frames. Figure 1(b) shows the plan view of the protype building. All steel columns were W12×106 shapes typically used for reaction frames in the NFRL and anchored to the laboratory strong floor. The perimeter columns were continuous over two stories, whereas the two interior columns were spliced 92 cm above the concrete slab on the first floor. The 9.1 m long W16×31 beams in the fire test bay were connected to the column flange or at midspan of the 6.1 m long W18×35 girder using standard shear tabs; see Figure 2(a). Extended shear tabs were used at the ends of W18×35 girders; see Figure 2(b). All other beams and girders were connected using bolted angles and extended shear tabs, respectively. All structural steel shapes and shear tabs were rolled from A992 steel (a minimum specified yield strength of 345 MPa) and A36 steel (a mimimum specified yield strength of 250 MPa), respectively.

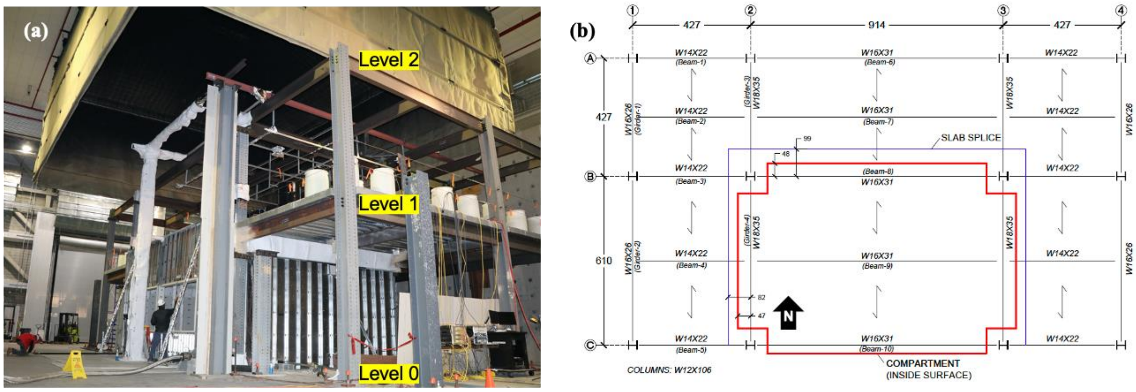


Figure 1. (a) Two-story prototype steel framed building; (b) Floor plan view (dimensions in cm).

The concrete slab was lightweight concrete (a minimum specified compressive strength of 28 MPa) cast on 76 mm deep formed steel decking. The concrete mixture included polypropylene fibres to minimize thermally induced spalling as suggested by Maluk at al. [4], which was also used for the Phase 1 composite beam study [5]. The concrete slab was cast approximately 5 months prior to fire testing. The moisture content of the concrete at time of fire testing was 7.6 % when measured according to ASTM C642-13 [6]. As shown in Figure 2(c), the slab thickness was 83 mm required for the 2-hour fire resistance rating with exposed steel deck. The cold-formed welded wire reinforcement of 59 mm^2/m, spaced at 152 mm, was embedded 41 mm below the top surface of the concrete slab. This is the minimum shrinkage reinforcement prescribed in the relevant U.S. design standard [7]. The floor slab was acting partially composite with steel beam assemblies through headed stud anchors (shaft Ø19 mm). For the W16×31 beams, a single headed

Choe, Lisa; Ramesh, Selvarajah; Dai, Xu; Hoehler, Matthew; Bundy, Matthew. "Experimental Study on Fire Resistance of a Full-Scale Composite Floor Assembly in a Two-Story Steel Framed Building." Presented at 11th International Conference on Structures in Fire (SiF 2020). November 30, 2020 - December 02, 2020.

stud anchor was spaced at 305 mm; a pair of the same size stud anchors were welded atop the W18×35 girders at spacing of 356 mm, as shown in Figure 2(c). The corresponding degree of composite action was about 65 % of the ambient yield strength of the steel beams. In addition, the 77 cm long No. 4 hooked reinforcing bars (Ø13 mm) were placed perpendicular to the south edge beam under each stud to prevent separation of the heated concrete slab during the test. The slab splices along the interior edges of the test bay, shown in Figure 1(b), incorporated No. 4 bars and screw anchors so that the surrounding floor assemblies can be reused for subsequent fire tests. This splice design allows the full slab continuity (i.e., moment and shear transfer) between the test floor and the surrounding floor assemblies at ambient temperature. The construction detail at the northwest corner of the test bay is shown in Figure 3(a).

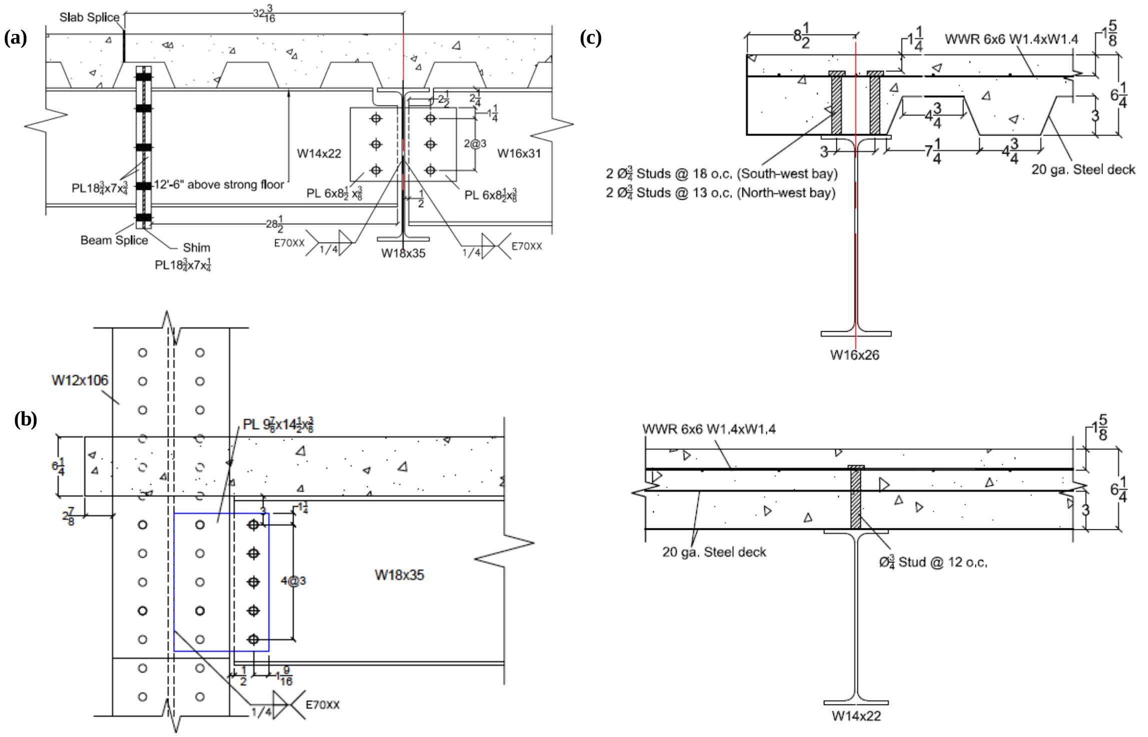


Figure 2. Details of (a) Standard shear-tab connection; (b) Extended shear-tab connection; (c) Composite sections. Dimensions in inch (1 in. = 2.54 cm)

Figure 3(b) shows a photograph of the fire test compartment, approximately 10 m long and 7 m wide. The height of the composite floor soffit was 377 cm above the compartment floor. The enclosing walls were constructed with 3 m tall sheet steel metals and a 48 mm thick gypsum board liner on the exposed wall surface. The 0.8 m gap between the compartment ceiling and the top edges of the three internal walls was lined with two layers of 25 mm thick ceramic blankets. Four natural gas burners (1.5 m × 1 m each) were distributed across the floor of the test compartment. The main vent was on the south wall, approximately 150 cm tall × 582 cm wide. There was a 30 cm tall × 582 cm wide slit on the north wall, designed for air intake only. The height of the windowsill on both the north and south walls was 100 cm above the strong floor.

The exposed steel frame within the test compartment, shown in Figure 1(b), was sprayed with a gypsum-based cementitious material (average specified density of 295 kg/m^3) for the 2-hour fire resistance rating. The average measured thickness of sprayed fireproofing prior to fire testing was 18 mm for both primary W16×31 beams and W18×35 girders and 13 mm for the secondary W16×31 beam. The coefficient of variation in thickness measurements was approximately 15 %. The exposed steel connections and columns were over sprayed with the same fireproofing material, with the thickness ranging from 25 mm to 28 mm.

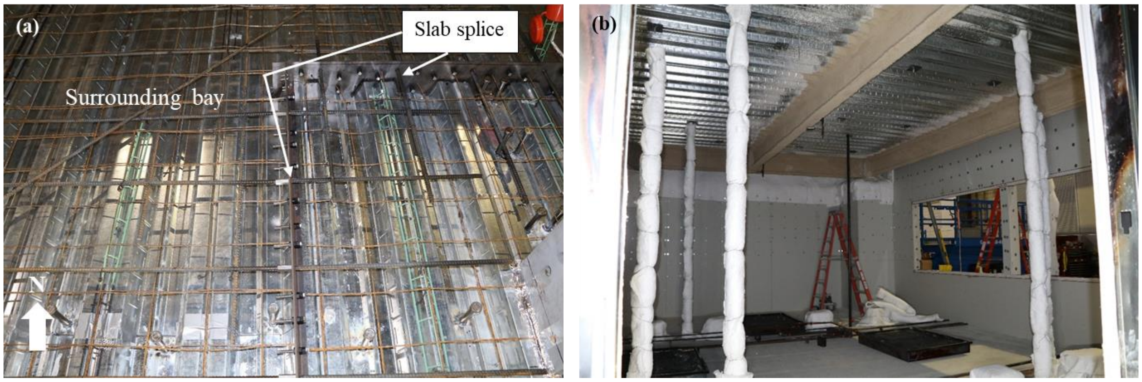
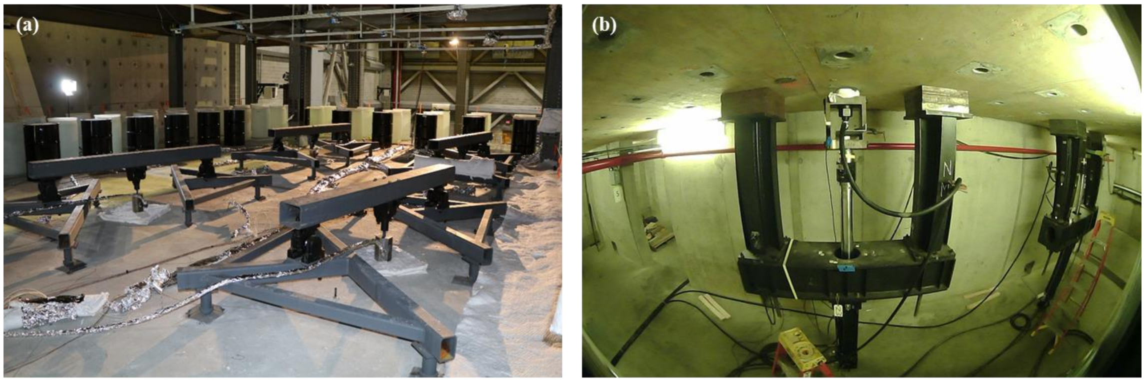


Figure 3. (a) Reinforcement details of slab splice; (b) Fire test compartment with the main vent on the south wall.

3 TEST CONDITIONS

The total mechanical load imposed on the 9.1 m × 6.1 m test floor assembly was approximately 150 kN (or 2.7 kPa), conforming to the American Society of Civil Engineers (ASCE) gravity load combination [8] of *1.2×dead load + 0.5×live load*. This load was distributed at twenty-four points across the test floor using water-cooled loading frames connected to four hydraulic actuators mounted in the basement, Figures 4(a) and 4(b). The total floor load including the assembly self-weight was approximately 5 kPa. The surrounding floors were loaded by water-filled drums, simulating a uniformly distributed mechanical load about 1.3 kPa. The load ratio (i.e., total load normalized by the ambient design capacity) of the test assembly was approximately 0.3 for both the secondary composite beam and the standard shear tabs of the same beam. The load ratio of the steel members and connections in the surrounding bays was less than 0.2.

The hydraulically loaded test floor assembly was exposed to a natural gas fuelled compartment fire, Figure 4(c), simulating the ASTM E119 time-temperature curve. Figure 4(d) shows the burner heat release rate (HRRburner) versus time relationship used in this test. This relationship was verified through a series of mock-up tests [10] conducted prior to this experiment. It should be noted that, as shown in Figure 4(d), there was a short loss (< 3 min) of the HRRburner data at 102 min due to network disruption of the natural gas delivery system. The fire and mechanical loads were removed at approximately 107 min. The total expanded uncertainty (with a coverage factor of 2 as defined in [11]) in measurements of the burner heat release rate and mechanical load is estimated 1.4 % at 10 MW [12] and 2 % at 150 kN, respectively.



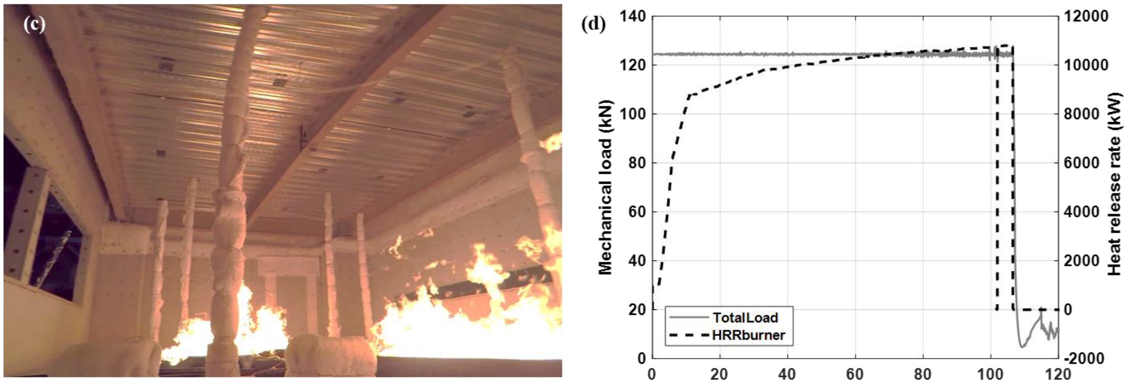


Figure 4. (a) Mechanical loading atop the test assembly; (b) hydraulic actuators mounted beneath the test compartment using yoke [9]; (c) compartment fire growth simulated with natural gas burners; (d) time history of total mechanical load (TotalLoad) and heat release rate of burners (HRRburner).

4 RESULTS AND OBSERVATIONS

4.1 Thermal response

Over two-hundred type K thermocouples (bead Ø 0.5 mm) were deployed at various locations across the test assembly. The average upper layer gas (ULG) temperature within the test compartment was measured using twelve Inconel-sheathed thermocouple probes hanging 305 mm below the steel deck. The average ULG temperature exceeded 700 °C at 11 min and reached a peak value of 1060 °C at 107 min. After 15 min from the burner ignition, the increase in the average ULG temperature resembled the International Organization for Standardization (ISO) standard 834 [13] temperature and was about 2 % higher than the ASTM E119 temperature. The standard deviation of temperatures measured using these thermocouples was less than 50 °C, indicating practically uniform temperatures below the test floor assembly.

Figure 5 illustrates the measured temperature rise in the midspan composite sections of the test assembly, Figure 2(c). For the W16×31 composite beams, the average temperature of the web and bottom flange of the protected steel beams reached 600 °C at 60 min and exceeded 800 °C at 107 min, Figure 5(a). The average concrete temperature at 0.5 mm above the top rib of the steel decking along the beam centrelines increased to 270 °C at 107 min. Temperatures of headed stud anchors at 38 mm above the top flange and welded wire reinforcement (WWR) remained below 400 °C and 200 °C, respectively.

For the W18×35 composite girders, as shown in Figure 5(b), the lower portion of the protected steel girders heated to 450 °C at 60 min and 700 °C at 107 min. The top flange steel temperature increased to 400 °C. The bottom concrete temperature in the shallow section next to the steel girder, Figure 2(c), increased to 600 °C at 80 min but significantly influenced by combined effects of concrete fractures and debonding of steel decking afterwards. The average temperatures of headed studs (at 2.5 cm above the steel decking) and WWR atop the girders never exceeded 300°C and 200 °C, respectively, during and after fire exposure.

The standard deviation in temperatures of the three heated W16×31 beams ranged from 60 °C to 110 °C, as indicated by error bars in Figure 5(a). This temperature variation might be caused by thermally induced fissures and degradation in coated insulation as the beams were undergoing severe thermal elongation and bending under fire loading. Unlike W16×31 beams, one can observe a smaller temperature difference (30 °C to 60 °C) between the east and west W18×35 girders, Figure 5(b). These girders seldom physically deformed, and thereby the applied fireproofing appeared to maintain relatively good integrity during fire loading.

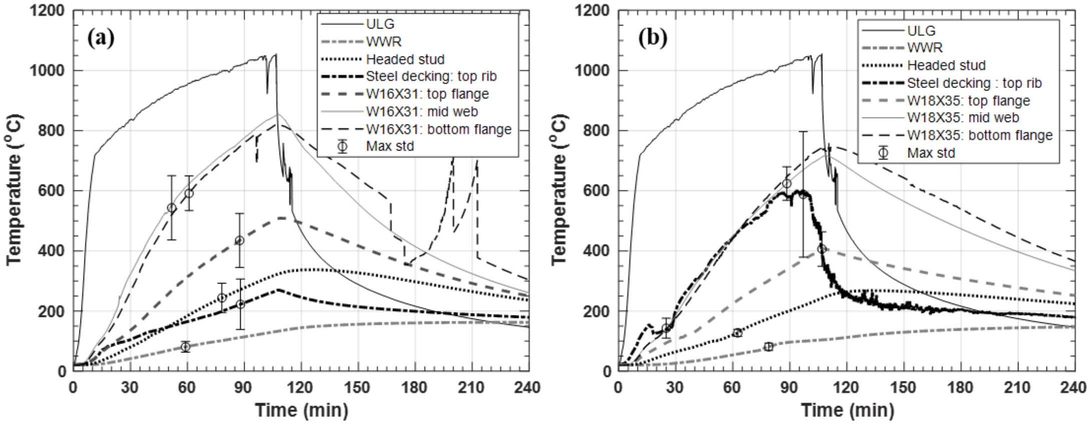


Figure 5. (a) Measured temperatures of (a) W16×31 composite beam and (b) W18×35 composite girder. Plotted with the average values of three W16×31 beams or two W18×35 girders in the test compartment.

A total of thirty-six thermocouples were mounted inside of the 9.1 m by 6.1 m test floor slab at various locations that were not thermally shaded by the steel framing underneath. Figure 6 shows the average concrete temperature in the deep (thickness = 159 mm) or shallow (thickness = 83 mm) sections of the concrete slab. Thermocouples installed at the steel decking (TST-5*) measured the hottest temperature, reaching nearly 900 °C during fire loading. The peak temperature of the concrete near the bottom rib of the steel decking (TST-4) was about 100 °C higher than the concrete temperature near the top rib (TST-7). However, the spatial temperature variation of TST-5* and TST-7 was quite high (> 110 °C), as indicated by the large error bars. These temperatures could be sensitive to debonding of the concrete from the steel decking. Temperatures of the WWR were affected by varying thickness of the concrete slab. At 107 min, for example, TST-1 (at deep sections) and TST-6 (at shallow sections) was 120 °C and 380 °C, respectively. In addition, the concrete temperatures towards the top surface (TST-1, TST-5, and TST-6) or at the centroid of the deep section (TST-2) were affected by moisture, as evidenced by the temperature plateau at 100 °C, over a longer period.

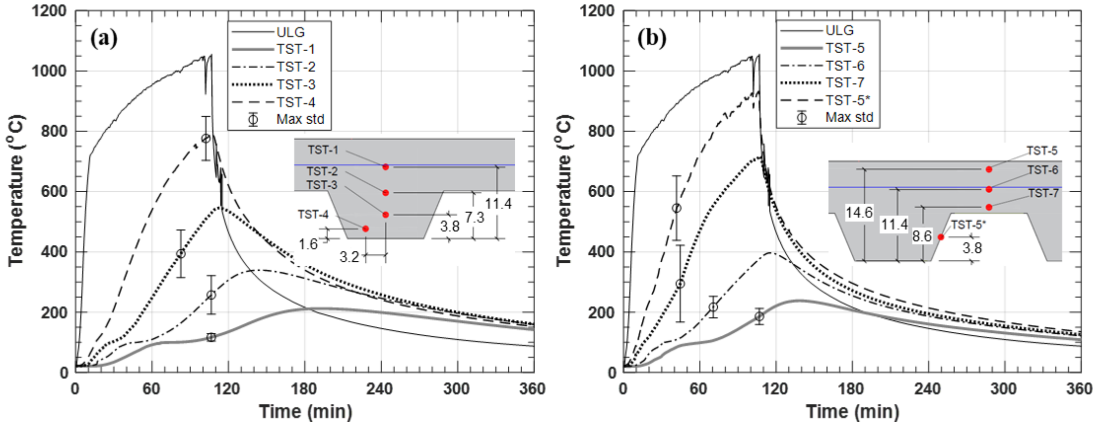


Figure 6. Measured temperatures of (a) deep section and (b) shallow section of the concrete floor slab. Plotted with the average values of six deep sections or four shallow sections across the test floor; thermocouple locations are in cm.

Although the data are not presented in this paper, the top (unexposed) surface temperature continued to rise during the cooling phase (up to 1 hour following extinguishment), ranging from 140 °C to 180 °C. Temperatures of the beam-to-girder shear-tab connection reached over 600 °C, whereas that of the shear-tab connections to columns were below 400 °C due to the thicker insulation sprayed on those regions. Detailed discussions and results of the connection temperatures are presented in the companion paper [14].

Estimates of total expanded uncertainty (with a coverage factor of 2) in measurements of the gas-phase, steel, and concrete temperatures are 8 % at 1110 °C, 4 % at 970 °C, and 6 % at 310 °C, respectively.

4.2 Structural response

The loaded test floor assembly continuously sagged during heating, while sequentially developing concrete fractures at various locations. No explosive spalling of the concrete was visible during or after the experiment, however, small 'popping' sounds continued during heating indicating that (micro) spalling was occurring between the bottom of the slab and the steel deck. Concrete surface cracks first appeared along the east and west girders as well as the north edge beam of the test bay about 20 min to 30 min after ignition. Around 40 min into heating, the southeast corner of the heated floor slab fractured making a loud noise. After 70 min, tensile fracture of the concrete was visible near the longitudinal (east-west) centreline of the test floor. Reaching 100 min in fire, small flames were intermittently visible above the top of the heated slab towards the east and west ends of this longitudinal crack, indicating failures of some screw joints of steel deck units in those locations. From this point forward, the mechanical loads on the south side of the test slab appeared to be supported by the steel deck and the south edge beam with concrete hanging cantilever, Figure 7(a). The fire and mechanical loading were removed at 107 min due to safety concerns.

A total of thirty displacement transducers were deployed to characterize the displacement of the two-story steel frame and the 9.1 m by 6.1 m test floor assembly during and after fire exposure. Figure 7(b) shows locations of the selected vertical and horizontal displacement sensors (labelled VD and HD, respectively) of the test assembly. All VD sensors in Figure 7(b) were located at the transverse (north-south) centreline of the test assembly. HD4 and HD6 sensors were used to measure thermal expansion at the perimeter of the heated floor assembly in the east-west direction and the north-south direction, respectively. HD9 measured the lateral displacement of the southeast column at the first story level. These horizontal displacement measurements were made at 15 cm above the top surface of the test floor slab. The total expanded uncertainty (with a coverage factor of 2) in measurements of the vertical and horizontal displacements is estimated 1 % at 580 mm and 5 % at 35 mm, respectively.

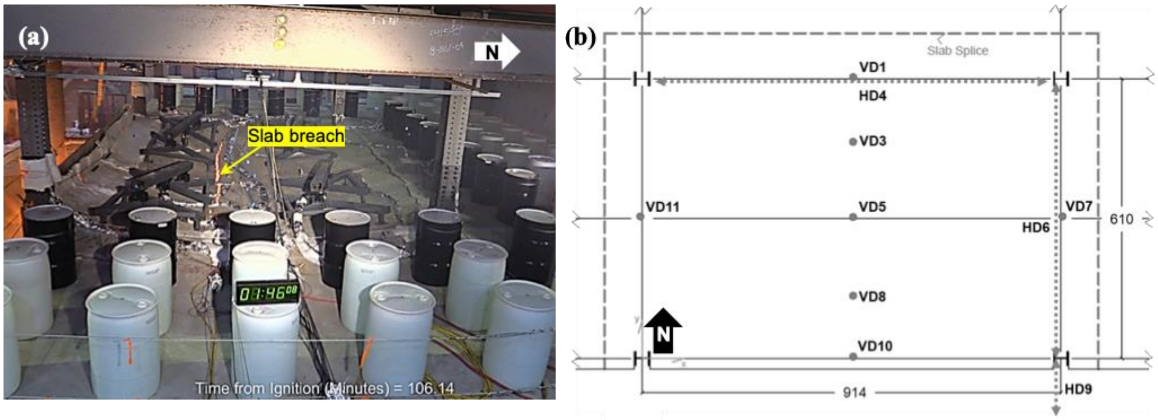


Figure 7. (a) Breach of the test floor slab; (b) Locations of displacement measurements with dimensions in cm.

As shown in Figure 8, the vertical displacement of the test floor assembly continuously increased during heating (until 107 min) and partly recovered during cooling; however, collapse did not occur. Until 60 min after ignition the values of VD5 and VD8 were similar. However, after the test floor slab began to breach (wide longitudinal crack) around 70 min, VD5 surpassed VD8 and reached 460 mm at 92 min. This displacement was approximately equal to the ratio of L/20 where L is the east-west span of 9.1 m. While VD5 finally reached the L/16 ratio at 107 min, there was no indication of 'runaway' deformation. Conversely, the vertical displacements of the perimeter steel members (VD1, VD7, VD10, and VD11) were relatively small, ranging from 65 mm to 210 mm. Also, VD7 and VD11 appeared to be less affected by the longitudinal concrete fractures. These perimeter members exhibited some degree of twisting and lateral deformations, discovered during the post-test inspections.

Choe, Lisa; Ramesh, Selvarajah; Dai, Xu; Hoehler, Matthew; Bundy, Matthew. "Experimental Study on Fire Resistance of a Full-Scale Composite Floor Assembly in a Two-Story Steel Framed Building." Presented at 11th International Conference on Structures in Fire (SiF 2020). November 30, 2020 - December 02, 2020.

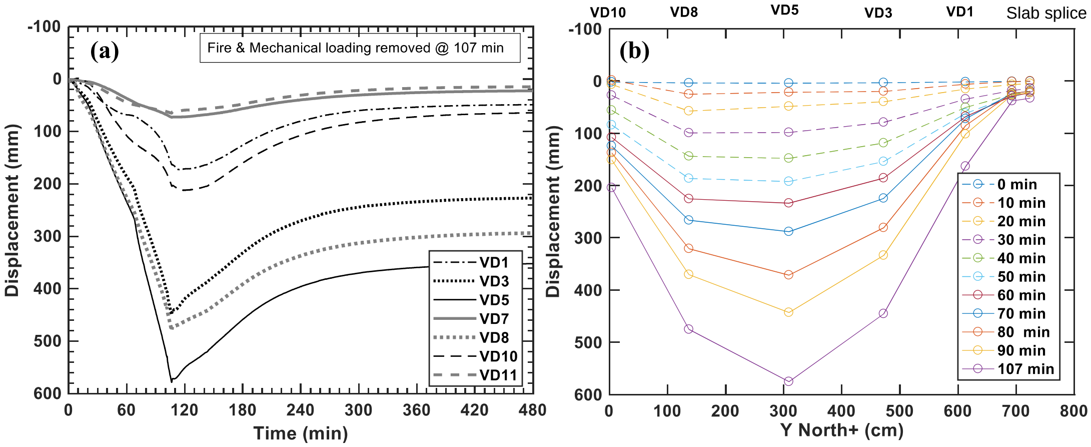


Figure 8. (a) Vertical displacements after the burner ignition at 0 min; (b) Vertical displacement profile varying with fire exposure time.

Figure 9(a) shows the midspan vertical displacements of fireproofing protected steel beams and girders as a function of the bottom flange temperatures. When the bottom flange temperature exceeded 700 °C, the vertical displacement of the secondary beam (VD5) increased much more rapidly from 0.4 mm/°C to 1.4 mm/°C. This change could be caused by several factors, such as initiation of a longitudinal breach of the test floor slab and continuous degradation of flexural strength and stiffness of support beams at higher temperatures. In the early stage of fire, on the other hand, the increase in displacements of the east and west girders (VD7 and VD11, respectively) was affected by smaller applied load ratios than the secondary beam. Furthermore, the heating rate of these members was relatively slow due to larger heat capacity and any heat loss associated with their close proximity to the upper wall lining with ceramic blankets or concrete fractures above these members. The vertical displacement of the south edge beam (VD10) was more responsive to the temperature change than other three perimeter members due to its free slab edge allowing less resistance to lateral-torsional buckling of this beam.

The horizontal (axial) displacements of the test assembly were measured using the lateral displacements of the columns at the first-story level. Figure 9(b) shows the time-varying horizontal displacement of the north primary beam (HD4) and the east girder (HD6) of the test assembly as well as the lateral displacement of the southeast column. The positive values in this figure represent the displacements due to thermal expansion of the heated test assembly. The values of HD6 and HD9 were similar throughout the test, indicating that the east girder expanded in single direction, towards the south due to a much larger restraint provided by the north surrounding frame. These displacements increased continuously to a peak value ranging from 32 mm to 34 mm until the fire test was terminated. The value of HD4 increased at a similar rate but began to decrease after 70 min when the longitudinal fracture of concrete occurred. The maximum axial displacement due to thermal expansion in the east-west direction was approximately 22 mm.

Figure 10 shows the final fracture pattern of the test floor slab after cooling. As mentioned earlier, concrete cracks developed along the north, east, and west edges of the test bay, followed by the longitudinal cracks 530 mm or less south of the secondary beam. Most of WWR (59 mm^2/m) across the thicker lines of fractures visible in Figure 10 completely ruptured. Neither concrete failures along the south (free) edge (i.e., separation from the south edge beam) nor slab splice failures were witnessed. Based on crack openings of the concrete, the east and west edge cracks were initiated near the flanges of the southeast and southwest columns, whereas the north edge crack was propagated from the midspan or its vicinity. No through-depth fractures were observed around the northeast and northwest columns.

The middle breach of the test floor slab appeared to be occurred due to catenary action in the north-south direction where the steel decking was continuous into the north adjacent bay. It is believed that the tensile membrane action of the test floor slab was not achieved or developed in a limited fashion because of the early formation of concrete fractures and ruptures of welded wire reinforcement along the east and west edges. These through-depth cracks located 100 mm or less inside of the test-bay column grid, which formed shortly after fire ignition and continued to widen unchecked during heating. Thus, the headed stud anchors on the east and west girders were ineffective to induce tension in the concrete in the east-west direction as the heated floor slab continued to deflect downward. In contrast, the north edge crack formed 370 mm or less north of the north primary beam (i.e., outside of the test-bay column grid) and thereby the headed stud anchors on all three 9.1 m long beams appeared to provide anchorage of the concrete and steel decking against tension developing in the north-south direction. As the vertical displacement of the test floor increased in fire, the excessive tension would develop more effectively in the north-south direction than in the east-west direction until WWR finally ruptured at critical locations. This WWR rupture would happen in the concrete where the vertical displacements were greater, i.e., south of the secondary beam as shown in Figure 8(b).

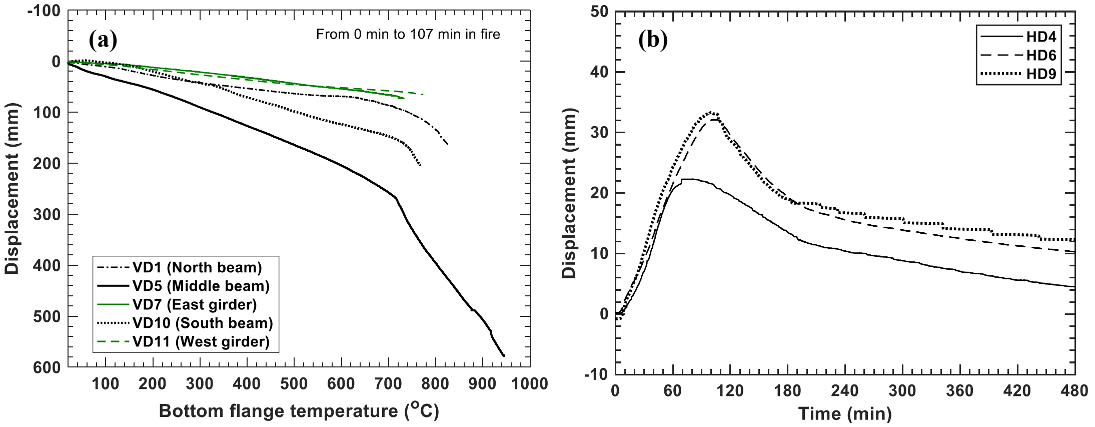


Figure 9. (a) Vertical displacement data as a function of bottom flange temperatures during heating; (b) Horizonal displacements measured at 15 cm above the test floor slab during heating (until 107 min) and cooling.

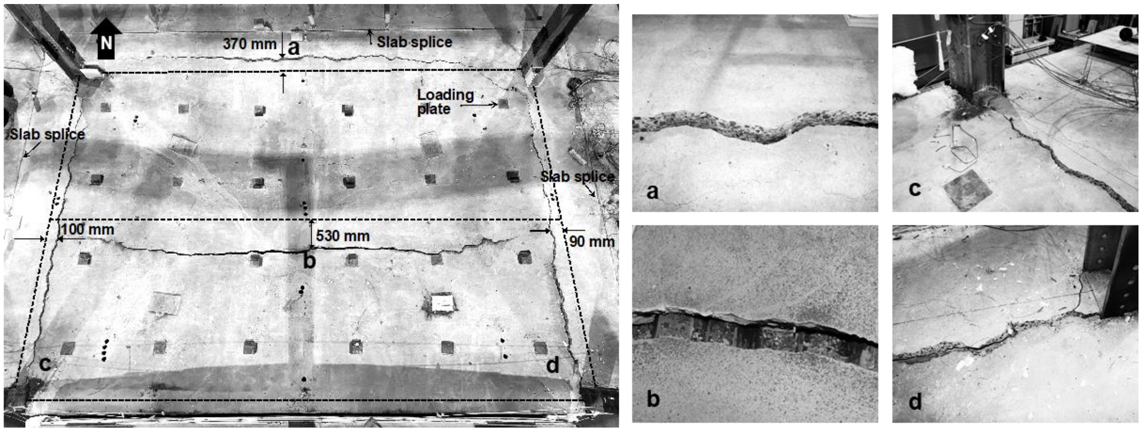


Figure 10. Top of the test floor slab after cooling. Dashed lines define the test-bay column grid. The four photographs on the right (a through d) show close-ups of concrete fractures.

Figure 11 shows the underside of the test floor assembly after cooling. The steel deck below concrete fractures (Figure 10) mostly maintained its integrity with good ductility. Only a local rupture was found in the deck unit below the east end of the mid-panel longitudinal crack, i.e., near the west edge of the top flange of the east girder. All three 9.1 m long W16×31 beams exhibited permanent strong axis bending deformation and local buckling near the beam ends. Furthermore, the north and south primary beams also exhibited twisting and lateral deformations. The end connections of these beams, however, maintained their structural integrity. The east and west W18×35 girders showed little residual vertical deflection and exhibited minor out-of-plane deformations in the webs near the end connections. The extended shear tabs

Choe, Lisa; Ramesh, Selvarajah; Dai, Xu; Hoehler, Matthew; Bundy, Matthew. "Experimental Study on Fire Resistance of a Full-Scale Composite Floor Assembly in a Two-Story Steel Framed Building." Presented at 11th International Conference on Structures in Fire (SiF 2020). November 30, 2020 - December 02, 2020.

welded to the northeast and the northwest columns exhibited noticeable out-of-plane deflection but no bolt failures. The extended shear tabs welded to the southeast and the southwest columns deflected little, whereas there were partial shear ruptures in the lower bolts of the southeast connection. The sprayed fireproofing on the beams and girders mostly remained intact, although fissures were evident on the beam web near the end connections and at the lower beam web of the secondary beam at midspan.

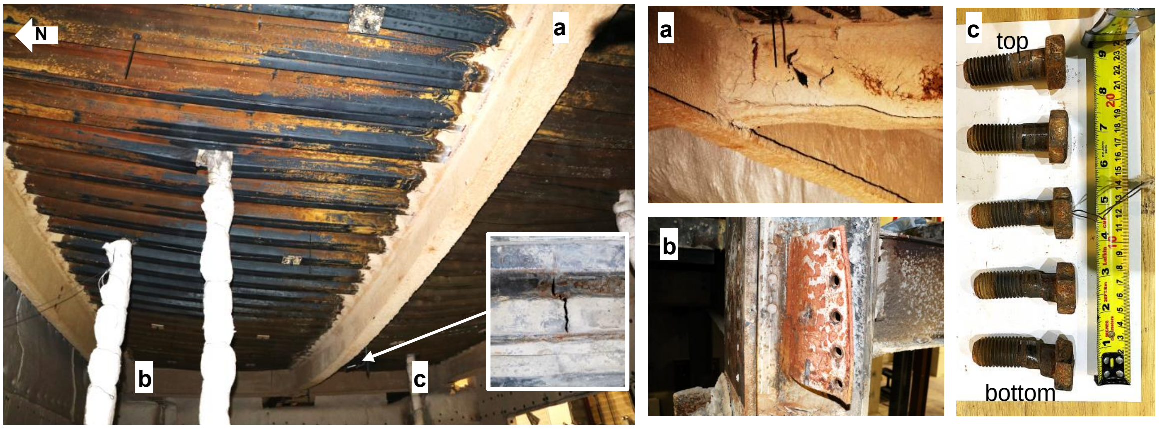


Figure 11. Fire-exposed steelwork of the test floor assembly after cooling. Close-ups of some deflected steel parts are shown in a through c.

4.3 Comparison with ASTM E119 criteria

The intent of standard fire testing, mostly performed using a purpose-built furnace, is to provide a consensus-based method to evaluate the duration for which an *isolated* floor assembly contains a fire while retaining its structural stability, the so-called *fire resistance rating* expressed in minutes or hours. This testing is typically performed using a test assembly with limited size (e.g., a minimum floor area of 16.7 m^2 and a minimum beam span of 3.7 m [3]) with two end support conditions, either *restrained* or *unrestrained* as explained in LaMalva et al [15]. A test assembly is required to resist its maximum load for normal conditions (e.g., *1.2×dead load + 1.6×live load* per ASTM E119 standard) while subjected to standard furnace heating. The fire resistance rating of a test assembly is usually determined based on limiting temperatures and displacements as discussed below.

Figure 12 summarizes the test results in comparison with ASTM E119 acceptance criteria. For the 2-hour *restrained* fire resistance rating, the test specimen must meet the following conditions: (i) sustaining the applied loads with no ignition of cotton waste placed on the top of the heated concrete slab during the full rating period, (ii) the average temperature on unexposed surface less than 139 °C above its initial temperature during the first hour, (iii) a peak temperature of structural steel members below 704 °C during the first hour, and (iv) the average temperature at any section of structural steel members below 593 °C during the first hour, and (v) the maximum total displacement less than the value of $Lc^2/400d$ where Lc = beam clear span, d = depth of composite beam, and the corresponding displacement rate less than the value of $Lc^2/9000d$. As shown in Figure 12, the protected individual W16×31 beams and W18×35 girders of the test assembly successfully met the limiting temperature and displacement criteria. The average concrete surface temperature measured by eight thermocouples distributed across the test assembly was approximately 120 °C prior to extinguishment of the fire. Although the maximum total displacement of the secondary beam exceeded the ASTM E119 displacement limit, the measured displacement rate of this beam was 40 % less than its specified value. It is important to note that this condition was achieved at the total floor load from the load combination of *1.2×dead load + 0.5×live load*, approximately 60 % of the maximum load condition (e.g., *1.2×dead load + 1.6×live load*) as prescribed in ASTM E119 [3].

Furthermore, this test has revealed some potential issues related to the integrity of a composite floor assembly as part of compartmentation under fire loading. As shown in Figure 13, the centre breach in the test floor slab, initiated prior to the specified rating period of 120 min, was accompanied by ruptures of the wire reinforcement in tension at the mid-panel displacement of 350 mm (L/26) or greater. A minimum code-required amount of shrinkage reinforcement (59 mm^2/m) used in the test assembly was insufficient to resist thermally induced tension during the investigated fire. Although the steel deck continuously running in the transverse (north-south) direction of the test assembly appeared to be ductile at large vertical displacements, failure of side deck joints (screws failure at the decking overlap), local deck ruptures, exposure of the heated decking units within concrete cracks allowed the penetration of flames and hot gases beyond the test compartment. This condition could have potentially ignited cotton waste placed on the unexposed surface, failing to meet the standard fire testing criterion (i) as mentioned above.

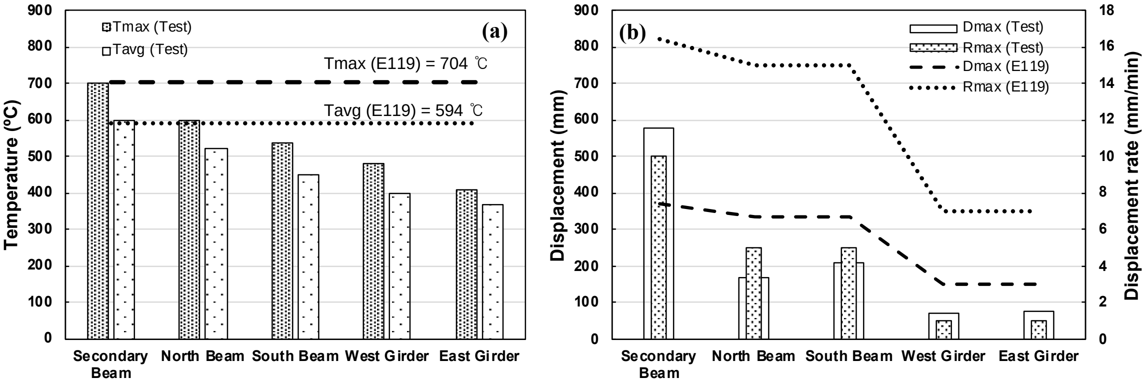


Figure 12. Comparisons of the test results with (a) limiting temperatures and (b) limiting displacements and displacement rates, where Dmax = maximum displacement and Rmax = maximum displacement rate.

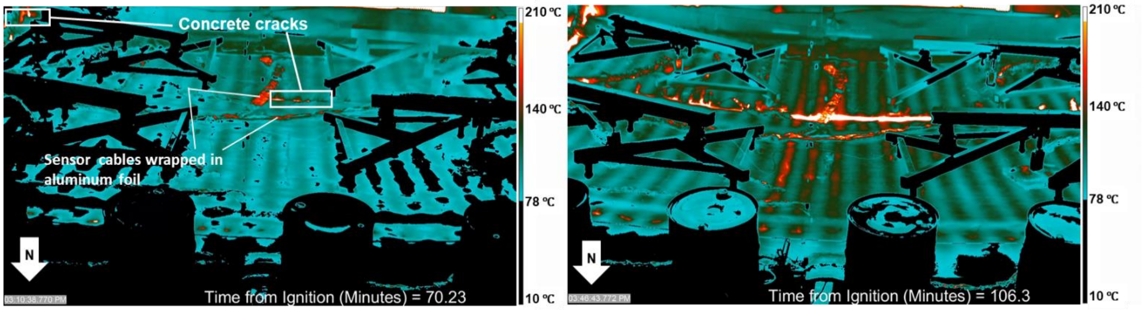

Concrete cracks
Sensor cables wrapped in aluminum foil
N
Time from Ignition (Minutes) = 70.23
210 ℃
140 ℃
78 ℃
10 ℃
N
Time from Ignition (Minutes) = 106.3
210 ℃
140 ℃
78 ℃
10 ℃

Figure 13. Thermal images of the top of the test floor slab at 70 min and 106 min after fire ignition.

5 SUMMARY & CONCLUSIONS

This paper presented the results of the first fire experiment on the 9.1 m by 6.1 m composite floor assembly situated on the first floor, south-edge bay of the two-story steel building designed and constructed following the current U.S. construction practice. The test floor assembly was subjected to a simulated compartment fire environment and mechanical loads conforming to the ASCE 7 load combination for extraordinary events (approximately 5 kPa including the assembly self-weight). The fire test conditions as well as thermal and structural responses of the test assembly to the combined effects of fire and mechanical loading are discussed and compared with the ASTM E119 acceptance criteria.

This test demonstrated that all fire-protected floor beams and girders met the ASTM E119 limiting temperatures. Also, these steel members never reached runaway at large vertical displacements (up to the ratio of L/16). The test floor assembly did not collapse after fire exposure, although some partial shear ruptures of connecting bolts were discovered after the test. However, the heated floor slab, during fire loading, exhibited a potential fire hazard before reaching a specified rating period, because of the use of the minimum code-compliant shrinkage reinforcement of 59 mm^2/m. The test floor slab began to crack along the interior edges (hogging moment regions) of the test column grid less than 30 min after ignition of the fire. The centre cracks appeared around the midspan of the secondary beam at 70 min, which continued to propagate in the east-west direction. The glowing hot deck was exposed on the top of the slab through enlarged concrete cracks. This main breach was caused by ruptures of wire reinforcement in tension (due to catenary action) parallel with formed steel decking. Membrane action of the floor slab appeared not to be effective due to ruptures of the wire reinforcement across the east and west edges of the test-bay column grid and the subsequent loss of the east and west vertical supports of the concrete slab. This initial experiment suggests that the minimum required slab reinforcement currently allowed in the U.S. practice may not be sufficient to maintain the structural integrity of the composite floor assembly during structurally significant fire events.

Further study is recommended to evaluate the fire hazard of relatively ‘thin’ concrete slab details permitted in steel building constructions. As future work, the second test in this experimental program will study the influence of enhanced slab design (for both strength and ductility) on the fire performance of composite floor systems and the effectiveness of tensile membrane action, which is believed to significantly improve the fire safety of steel-framed buildings. The influence of slab reinforcement on the fire performance of composite floor systems is discussed in Choe et al. [16], which is the basis of the experimental design of the second test being planned in 2021.

ACKNOWLEDGMENT

This work was conducted as part of the project “Measurement of Structural Performance in Fire” under the NIST Engineering Laboratory’s Fire Risk Reduction in Building Programs. The authors thank William Baker (Skidmore, Owings, and Merrill), Craig Beyler (Jensen Hughes), Luke Bisby (University of Edinburgh), Ian Burgess (University of Sheffield), Charles Carter (AISC), Charles Clifton (University of Auckland), Michael Engelhardt (University of Texas), Graeme Flint (Arup), Nestor Iwankiw (Jensen Hughes), Kevin LaMalva (Warringtonfire), Roberto Leon (Virginia Tech.), and Amit Varma (Purdue University) for their expert consultation. The authors also thank the NIST colleagues including Brian Story, Anthony Chakalis, Philip Deardorff, Laurean DeLauter, Marco Fernandez, Artur Chernovsky, Rodney Bryant, William Grosshandler, John Gross, Mina Seif, Ana Sauca, Fahim Sadek, Joseph Main, Chao Zhang, and Jonathan Weigand for their significant contributions to design, construction, and execution of this test program.

REFERENCES

1. Choe, L., Ramesh, S., Grosshandler, et al. Composite floor beams with simple shear connections subject to compartment fires: experimental evaluation, Journal of Structural Engineering, vol. 146, pp. 1-14, 2020. https://doi.org/10.1061/(ASCE)ST.1943-541X.0002627
2. Bundy M., Hamins, A., Gross, J., Grosshandler, W., Choe, L., Structural fire experimental capabilities at the NIST national fire research laboratory, Fire Technology, vol. 52 (4), pp. 959-966, 2016. https://doi.org/10.1007/s10694-015-0544-4
3. ASTM, Standard methods of fire test of building construction and materials, ASTM E119−19, ASTM International, West Conshohocken, PA, 2019.
4. Maluk, C., Bisby, L., Terrasi, G.P., Effects of polypropylene fibre type and dose on the propensity for heat-induced concrete spalling, Eng. Struct., vol. 141, pp. 584–595, 2017. https://doi.org/10.1016/j.engstruct.2017.03.058
5. Ramesh, S., Choe, L., Seif, M., et al. Compartment fire experiments on long-span composite beams with simple shear connections Part 1: experimental design and beam behavior at ambient temperature, Technical Note (NIST TN) - 2054, p141, 2019. https://dx.doi.org/10.6028/NIST.TN.2054
6. ASTM, Standard Test Method for Density, Absorption, and Voids in Hardened Concrete, ASTM C642 - 13, ASTM International, West Conshohocken, PA, 2013. https://doi.org/10.1520/C0642-13
7. SDI, C-2017 Standard for composite steel floor deck-slabs, Steel Deck Institute (SDI), 2017. http://www.sdi.org/wp-content/uploads/2017/02/ANSI-SDI-C-2017-Standard.pdf
8. ASCE, Minimum Design Loads and Associated Criteria for Buildings and Other Structures, ASCE/SEI 7-16, American Society of Civil Engineers, Reston, VA, 2016.

9. Choe, L., Ramesh, S., Hoehler, M. National fire research laboratory commissioning project: testing steel beams under localized fire exposure, Technical Note (NIST TN) – 1983, p117, 2018. https://dx.doi.org/10.6028/NIST.TN.1983
10. Zhang, C., Grosshandler W., Sauca A., and Choe L. Design of an ASTM E119 fire environment in a large compartment, Fire Technology, pp. 1–23, 2019. https://doi.org/10.1007/s10694-019-00924-7
11. Taylor B. and Kuyatt, C. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, Technical Note (NIST TN) – 1297, p24, 1994. https://doi.org/10.6028/NIST.TN.1297
12. Bryant, R. and Bundy, M. The NIST 20 MW calorimetry measurement system for large-fire research, Technical Note (NIST TN) – 2077, p68, 2019. https://doi.org/10.6028/NIST.TN.2077
13. ISO, ISO 834-2:2019 Fire-Resistance Tests – Elements of Building Construction (ISO, Geneva), 2019. Available at https://www.iso.org/standard/75137.html
14. Dai, X., Choe, L., Fischer, E., Clifton, C. Thermal response and capacity of beam and shear connections during a large compartment fire experiment. Proceeding of the 11th International Conference on Structures in Fire (SiF' 20), November 30 – December 02, 2020, University of Queensland, Australia (accepted for publication)
15. LaMalva, K., Bisby, L., Gales, J. et al. Rectification of "restrained vs unrestrained". Fire and Materials, pp 1-11, 2020. https://doi.rog/10.1002/fam.2771
16. Choe, L., Ramesh, S., Zhang, C., Clifton, C. Behaviour of composite floor assemblies subject to fire: Influence of slab reinforcement. Proceeding of 2021 Eurosteel Conference, September 1-3, 2021, University of Sheffield, United Kingdom (in production)

Behavior of Composite Floor Assemblies Subject to Fire: Influence of Slab Reinforcement

Lisa Choe[1], Selvarajah Ramesh[1], Chao Zhang[1], Charles Clifton[2]

Correspondence

Dr. Lisa Choe
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899, USA
Email: lisa.choe@nist.gov

Abstract

A series of fire experiments are being conducted on the 2-story steel framed building with composite floors in the National Fire Research Laboratory. This paper presents a brief overview of experimental design and discusses some results of the first test conducted on 6.1 m by 9.1 m composite floor assembly exposed to a standard fire. The test floor slab was constructed with lightweight concrete cast on 7.6 cm deep formed steel deck units and welded wire fabric (60 mm/m^2) as the minimum shrinkage reinforcement required in the U.S. practice. The structural steel beams and girders were sprayed with a cementitious fire resistive material for a 2-hour restrained fire resistance rating. While the test floor assembly resisted mechanical loads (2.7 kN/m^2) applied using hydraulic actuators, its underside was exposed to a natural gas-fueled compartment fire that was equivalent to the standard gas temperature-time relationship. The study revealed that the test floor assembly could withstand an imposed load at a large vertical deflection on the order of span / 16, but the concrete slab failed due to limited ductility and strength prior to 2 hours in a standard fire. Simple post-test predictions were performed incorporating tensile membrane action in the floor slab to compare with the measured behavior. Limited comparisons discuss the contribution of steel decking and slab reinforcement to the load-carrying capacity of a floor assembly in fire.

Keywords

Composite floors, compartment fires, standard fires, fire resistance

1 Introduction

Fire safety design of steel-framed buildings in the United States is based on prescriptive fire-resistance ratings of individual load-bearing elements with insulation. In real fires, however, the actual fire resistance of composite floor assemblies can be largely influenced by restraints and stiffness provided by adjoining structures that often remain cool. If structural redundancy of a given composite floor assembly against fire attack is unknown, fire ratings mandated in building codes can result in high installation cost of passive fire protection systems in multistory buildings but do not necessarily guarantee increasing fire safety.

Over the last few decades, there have been active experimental studies to assess the fire resistance of composite steel frames in Europe. The Cardington fire test program [1–2], in particular, highlighted that the actual fire resistance of composite floor assemblies in a steel-framed building surpassed that of isolated floor assemblies used for standard fire testing. With the secondary load-carrying mechanisms, i.e., tensile membrane action, composite floor assemblies supported by primary structural steel frames can withstand the fire loading without collapse, thus secondary floor beams can remain unprotected. The load-carrying capability of composite floor systems observed in those tests was highly influenced by structural steel connections and steel reinforcement used in composite slabs. Connection details and high reinforcement ratios used in those tests are more acceptable in the European construction practice.

Although the methods for high-fidelity modeling have evolved, they still require validation with test data with quantified uncertainty in measurements. To date, there is lack of data describing the actual fire performance of full-scale composite steel frames designed in compliance with the United States building codes and specifications. Experimental tests studying the full-scale span and size of floor assemblies commonly used in common construction practice cannot be achieved using the standard testing method with furnaces.

Motivated by such research needs, a multi-year research project is being conducted at the National Institute of Standards and Technology (NIST) to study the fire behavior and design limit states of full-scale composite floor systems. The experiments conducted at

[1] National Institute of Standards and Technology, Gaithersburg, USA
[2] University of Auckland, Auckland, New Zealand

Choe, Lisa; Ramesh, Selvarajah; Zhang, Chao; Clifton, Charles. "Behavior of Composite Floor Assemblies Subject to Fire: Influence of Slab Reinforcement." Presented at Eurosteel Sheffield 2021 (9th European Conference on Steel and Composite Structures), Sheffield, UK. September 01, 2021 - September 03, 2021.

NIST represent a major advance in full-scale experimentation of steel-concrete composite floor systems under fire and structural loading. The proposed test series capture a broad spectrum of geometric, design, and loading parameters relevant to current construction practice. This paper presents an overview of the first test of this program, design of the test fire, and some results of the test floor assembly in fire.

2 Test structure

Figure 1 shows a 3D rendering of the two-story steel-framed building constructed at the NIST National Fire Research Laboratory [3]. This 7.2 m tall prototype building has two bays by three bays floor plan which covers a nominal area of 18 m × 11 m. The fire test bay was located at the middle edge bay that was 6.1 m by 9.1 m in plan and 3.8 m in height above the strong floor. In this way, other surrounding bays acted as restraints to the test floor assembly while the continuity of the test floor slab was achieved. The second-story steel framing was to simulate the column continuous over two stories. Also, slab splices were designed to enable replacement of the test floor assembly new for subsequent tests. During the test, the slab continuity was acheived through steel reinforcement (No. 4 reinforcing bars) and shear connectors in the region of slab splices.

The test structure was designed according to the U.S. building codes and design specifications (e.g., [4–6]) to resist a construction live load of 0.96 kN/m^2, super imposed dead load of 0.48 kN/m^2, and a live load of 3.35 kN/m^2 at ambient temperatures, using the load combination of 1.2×dead load + 1.6×live load.

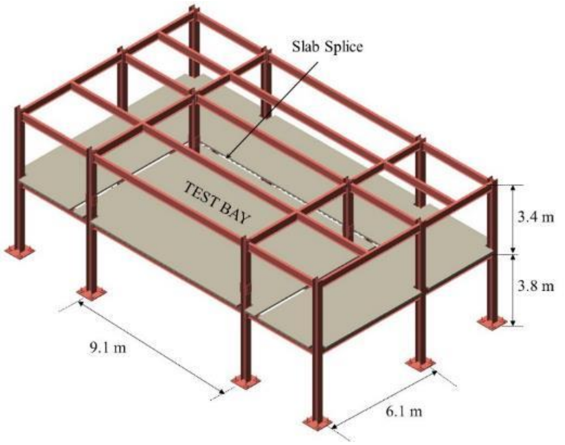


Figure 1 Test Structure-3D view

2.1 Structural frame

The structural steel gravity frames commonly used in the office buildings were designed to support composite floors. Figure 2 shows the layout of steel framing. In the test bay, a W16×31 shape was selected for three 9.1 m long beams, and a W18×35 shape was selected for the two 6.1 m long girders based on the serviceability criteria of floor vibration. All steel beams and girders were made of ASTM A992 [7] steel and were not cambered. The W12x106 shapes that were available in the laboratory as an erector set were used for columns and they were anchored to the strong floor using high strength post-tensioned bars.

Single-plate connections (shear tabs) were used for the test floor beam or girder connections. For the end connection of W16x31 beams (Figure 3), a 10 mm thick plate (ASTM A36 [8]) was bolted to the beam web using three 19 mm diameter bolts (Gr. A325 specified in the ASTM F3125 [9]) and welded either to the girder web or to a sacrificial plate on the column flange using a 6 mm fillet weld. For the girder-to-column connection (Figure 4), a 10 mm thick extended shear tab was bolted to the girder web using five 19 mm diameter bolts and welded to the web of the W12×106 column using a 6 mm fillet weld. Standard short-slot holes (21 mm in width and 25 mm in length) were used in shear tabs to allow the erection tolerances of ± 3 mm.

Sprayed fire resistive materials (SFRM) were applied to steel members exposed to fire. A medium density (ranging from 240 kg/m^3 to 350 kg/m^3) gypsum-based cementitious material was used. For a 2-hour fire resistance rating, the SFRM thickness was 17 mm for both primary beams and girders and 11 mm for the secondary beam.

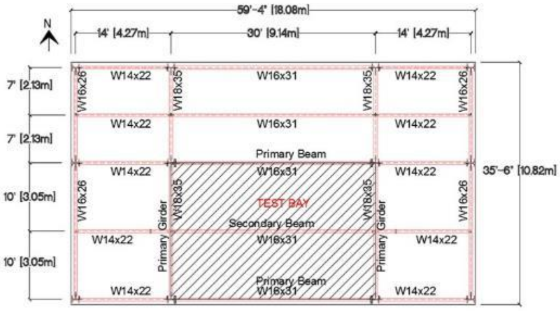


Figure 2 Steel frame layout-plan view

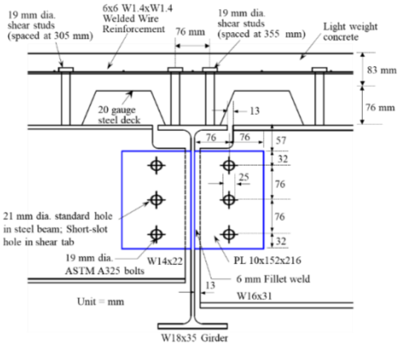

19 mm dia. shear studs (spaced at 305 mm)
6x6 W1.4xW1.4 Welded Wire Reinforcement
76 mm
19 mm dia. shear studs (spaced at 355 mm)
Light weight concrete
83 mm
76 mm
20 gauge steel deck
13
76
76
57
32
25
76
76
32
21 mm dia. standard hole in steel beam; Short-slot hole in shear tab
19 mm dia. ASTM A325 bolts
W14x22
PL 10x152x216
6 mm Fillet weld
13
W16x31
Unit = mm
W18x35 Girder

Figure 3 Shear tab connection between beam and girder

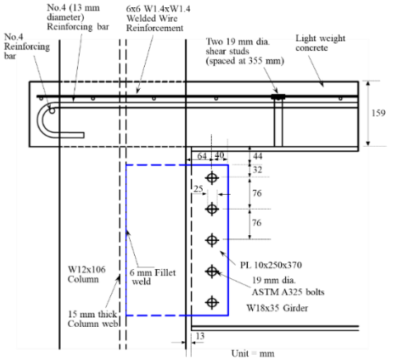

No.4 (13 mm diameter) Reinforcing bar
6x6 W1.4xW1.4 Welded Wire Reinforcement
No.4 Reinforcing bar
Two 19 mm dia. shear studs (spaced at 355 mm)
Light weight concrete
159
64
40
44
32
25
76
76
W12x106 Column
6 mm Fillet weld
PL 10x250x370
19 mm dia. ASTM A325 bolts
W18x35 Girder
15 mm thick Column web
13
Unit = mm

Figure 4 Extended shear tab connection between girder and perimeter column

2.2 Composite floor

The composite floor consisted of lightweight concrete slab cast on a 76 mm deep formed steel decking. Monofilament polypropylene microfibers, in the amount of 2.37 kg/m^3, were used in the concrete mix to reduce the likelihood of spalling [10]. An 83 mm thick slab above the steel deck was selected for a 2-hour fire rating with exposed steel deck. The ribs of the steel deck were oriented perpendicular to the W16x31 beams and parallel to the W18x35 girders. Figures 3 and 4 also illustrate the concrete slab details above beams and girders. Steel headed stud anchors (ASTM A29 [11]) with the diameter of 19 mm provided composite action of approximately 65% between the concrete slab and W16×31 beams or W18×35 girders.

Welded wire fabric (3.4 mm diameter cold formed plain steel wires spaced 150 mm) were placed at the mid-height of the topping concrete as required minimum area of shrinkage reinforcement specified in the design specification [6].

No.4 reinforcing bars (minimum yield strength of 414 MPa) with 180-degree hooks were placed below the heads of stud anchors welded on the edge beam to prevent premature failure in concrete slab due to lack of continuity in the south edge of the test bay.

2.3 Gravity loading setup

A total mechanical load of 125 kN was applied to the test floor assembly during the fire test. When combined with the weight of the test floor assembly and loading system, this load level was equal to 5.1 kN/m^2, the design load combination [4] with fire events (1.2×dead load + 0.5×live load). The applied load was equivalent to 20−30% of the ambient flexural capacity of the composite floor beams and 20−40% of the ambient shear capacity of shear connections.

Four hydraulic actuators mounted (in the basement) below the fire test bay were used to apply uniform loads on the test floor assembly. As shown in Figure 5, the actuator loads were transferred to purpose-built loading systems placed on top of the test floor assembly via water-cooled structural steel tubes running through the fire test bay and the strong floor. The applied loads were distributed at 24 points over the 6.1 m by 9.1 m test floor. A total of seventy-six water filled drums (2.1 kN each) were uniformly distributed over the surrounding floors, providing a gravity load of 1.2 kN/ m^2 (equivalent to 0*.5*×live load).

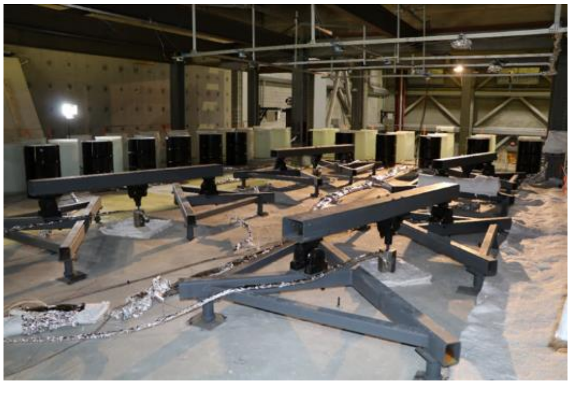


Figure 5 Photograph of the test floor **Source: NIST**

3 Test fire

3.1 Design objective

Standard fire testing [12] is aimed to develop uniform gas temperatures within a furnace by following the prescribed time-temperature curve. However, the size of specimens is often limited by that of furnaces available in fire testing facilities. In this test program, thus, the natural gas fuel delivery system [3] was used instead.

The key elements considered for design of the test fire include room geometry, ventilation, and fuel load. The fire exposure to the test floor assembly is designed to (i) be confined within a compartment, allow flame leakage through openings with restricted sizes and locations; (ii) produce the uniform gas temperatures in the upper layer of the compartment below the test floor assembly; and (iii) be controllable and repeatable with the use of existing natural gas fueled burners in the lab.

3.2 Fire confinement & burners

In order to confine a fire below the test floor assembly, the fire test compartment (10 m x 6.9 m x 3.8 m) was constructed with 3 m tall light-gauge steel walls protected by 50 mm thick type-C gypsum boards (Figure 6a). The upper portion of the test compartment (0.8 m in height) was confined by 50 mm thick ceramic blankets in order to allow deflection of the floor assembly without damaging the compartment wall during the fire test or generating restraint from the wall. The main ventilation opening (6 m wide, 1.5 m high) was in the south wall, while the slit on the opposite (north) wall was designed for air intake only. Four 1 m by 1.5 m natural gas burners, rated up to 4MW each, were distributed on the floor of the test compartment (Figure 5). The test fire conditions, ventilation openings and location of burners were designed using numerical simulations and mockup fire tests presented in Zhang et al [13] and briefly summarized herein.

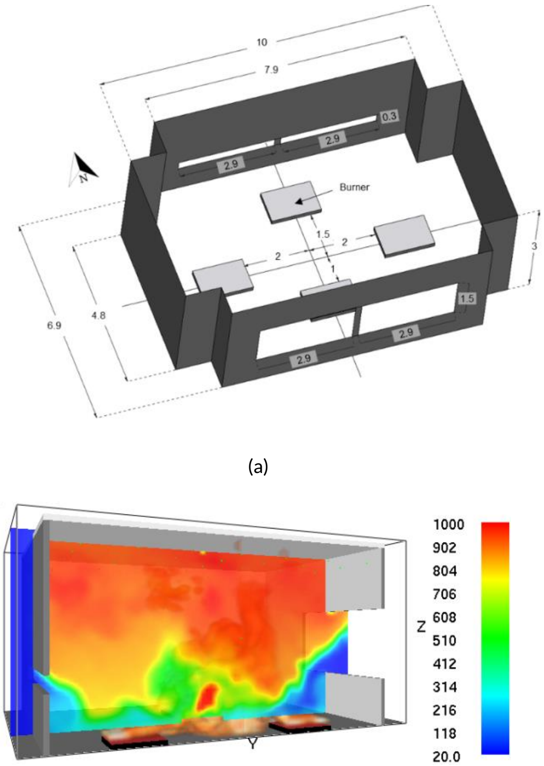
3.2.1 Fire Load

The maximum value of heat release rate (HRR) considered in this study was determined based upon knowledge gained in previous full-scale experiments [2, 14–15] with a similar compartment size. Surveys [Vassart, 2014] have found that the fuel loads in commercial and public spaces vary greatly. A typical office contains in the range of 420 to 655 MJ/m^2 of combustible material; a shopping center is in the range of 600 to 936 MJ/m^2; and a library can have fuel loads up to 2340 MJ/m^2. Thus, the natural gas burners used in this test program was designed to create a fuel load density (q_f)of approximately 1200 MJ/m^2 for a two-hour or longer fire exposure to attain significant structural failure. The peak intensity of the fire on a volumetric basis is 38 kW/m^3.

3.2.2 Opening factor

The test fire considered in this study is designed to maximize the upper layer temperature, to minimize the level of smoke, and to avoid excess fuel feeding a fire external to the bay. The ventilation is controlled by the total opening area, A_o, and the height of the opening, H_o. When scaled with the room volume, the opening area in this study, 0.034/m, is similar to the opening area/volume (A_o/V) used in the over-ventilated Cardington fire. The corresponding opening factor (F_o) is 0.045 m$^{1/2}$ where $F_o = A_o H_o^{1/2} A_t^{-1}$ and A_t(= 70 m^2) is the area of internal compartment boundaries including openings.

Once the initial values of fire load density (q_f) and opening factor (F_o) for the test fire were chosen, a zone model, CFAST [16], was used to compute the upper layer gas temperature varied with characteristics of openings (quantity, geometry, and distribution). With the opening geometry shown in Figure 6, CFAST results showed that the proposed *HRR*, 10MW was sufficient to produce the uniform upper layer temperature in the range of 1000 °C to 1200 °C. In addition, a computational fluid dynamics (CFD) model developed using Fire Dynamics Simulator (FDS) [17] indicated that with the proposed opening factor of 0.045 m$^{1/2}$, a fire appeared to be over-ventilated while well confined in the compartment. The predicted gas temperatures were deemed uniform (Figure 6b).



(a)

(b)

Figure 6 (a) Test compartment geometry and location of burners (unit = m), (b) predicted gas temperatures (unit = °C) using FDS model [18]

4 Results and Discussion

Almost 500 data channels were used to characterize the imposed fire and loading conditions as well as thermal and structural responses of the test floor assembly. This paper only focuses on gas temperature measured in the upper layer of the test compartment, the average temperatures in the test floor slab and the secondary beam, and slab deflections at selected locations. A full test report which contains the entire test data will be published in 2020. Refer to NIST Reports [10, 19] for the uncertainty in measurements reported in this paper.

4.1 Upper layer gas temperature

Figure 7 shows the HRR measured at the burner and the average upper layer gas temperature measured at 30.5 cm below the exposed surface of the test floor. The average upper layer gas temperature reached 950 °C at 60 min and a peak value of 1050 °C when the fire was extinguished at 107 min. At about 102 min, the fire was temporarily disrupted due to a network problem of data acquision system. As shown, measured upper layer gas temperatures were compared well with those predicted using FDS as well as the ASTM E119 [12] and ISO 834 [19] time-temperature curves.

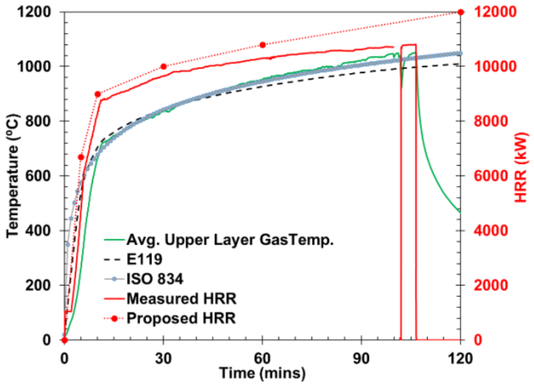


Figure 7 Upper layer gas temperature and burner HRR

4.2 Thermal response

Figure 8 shows locations of the embedded thermocouples used to measure the concrete temperature. Figure 9 shows the average temperatures measured at various depths within the concrete. The bottom concrete reached 700 °C to 800 °C at 107 mins. The error bars on these graphs represent the maximum standard deviation of measured temperatures across the test floor during heating and cooling phases. The temperature at mid depth of the topping concrete, where the welded wire fabric was located, varied due to the difference in concrete mass between the deep and the shallow sections. The temperature difference between these two sections was as high as 200 °C during the fire exposure. The temperature of the upper concrete continued to increase after the fire was extinguished. Figure 10 shows the temperature of the SFRM-coated secondary beam at midspan. Total expanded uncertainty (with a 95 % confidence interval) in measurements of the gas-phase, steel, and concrete temperatures are 8 % at 1110 °C, 4 % at 970 °C, and 6 % at 310 °C, respectively.

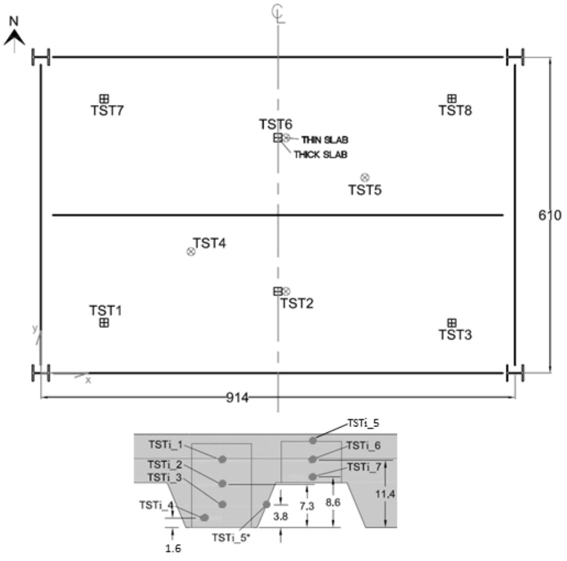
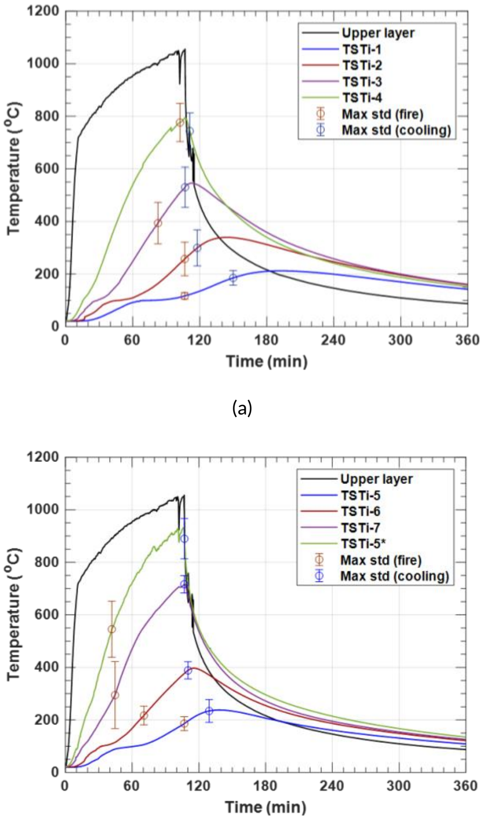


Figure 8 Locations of embedded thermocouples in the concrete slab (unit = cm)



(a)

(b)

Figure 9 Concrete slab temperature (a) through deep section, (b) through shallow section

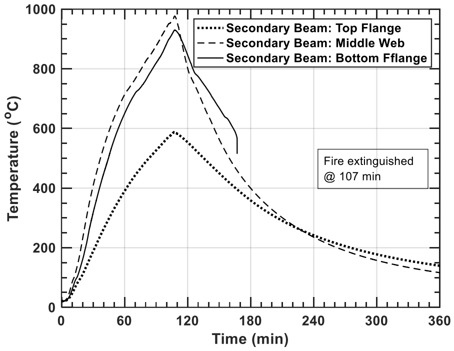


Figure 10 Temperature of the secondary (W16x31) beam at midspan

4.3 Slab deflection

Figure 11 shows locations of the vertical displacement measurements of the slab and Figure 12 shows the vertical displacements measured during the heating and cooling phases. Throughout the fire test, the south side of the slab (VD8 and VD10) deflected more than the north side (VD1 and VD3) due to the absence of the slab continuity in the south side. The displacements measured at VD5 was similar to that of VD8 until 60 min and became greater afterwards possibly due to local buckling at the end of the secondary beam and concrete fracture along the west and east girders. The mid-panel deflection (VD5) increased to 460 mm (equivalent to the ratio of L/20 where L = 9.1 m) at 92 min and to its peak value of 580 mm (L/16) without collapse when the fire was extinguished around 107 min. The total expanded uncertainty (with a 95 % confidence interval) in measurements of the vertical displacements is estimated 1 % at 580 mm.

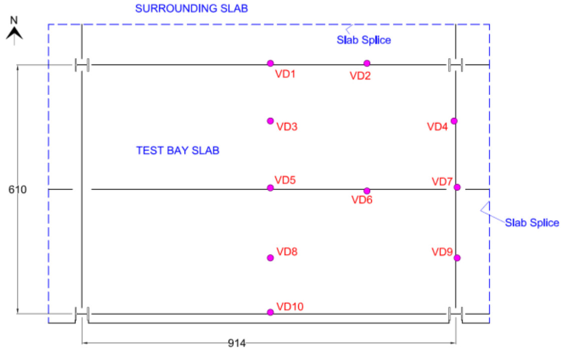


Figure 11 Location of vertical displacement measurements on the concrete slab

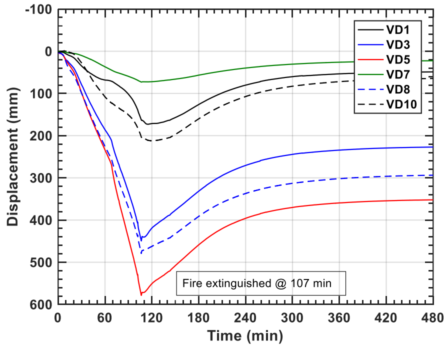


Figure 12 Vertical displacement of the test floor slab along the slab centrelines

4.4 Concrete fracture failure

Figure 13 shows photographs of the test floor slab after cooling, including concrete fracture along the perimeter of the test bay (west and east girders as well as the north primary beam) and along the secondary beam. Before reaching 30 min in fire, concrete cracked along the perimeter of the test bay. The cracks along the west and east side of the test floor were formed in the shallow section of the trapezoidal deck in the test bay next to the 6.1 m long girders. This is the critical section subjected to a large vertical shear and a hogging moment due to the mechanical loads on the test floor. The north longitudinal cracking was developed outside of the column grid. This cracking was developed between the beam centerline and the slab-splice plate. There was no crack along the south beam which had a free slab edge. Most of welded wire reinforcement across the perimeter cracks ruptured except for the interior column regions. Neither concrete cracking nor ruptures of reinforcement was observed in the slab splices used in this experiment.

Along the secondary beam, tensile concrete fracture started developing around the midspan at approximately 70 min and propagated towards the west and east directions. The steel deck locally fractured near the east end of this longitudinal crack (Figure 13), and small flames were visible on top of the slab until the fire was extinguished at 107 min. No collapse failure occurred, whereas most of welded wire reinforcement in the longitudinal crack ruptured.

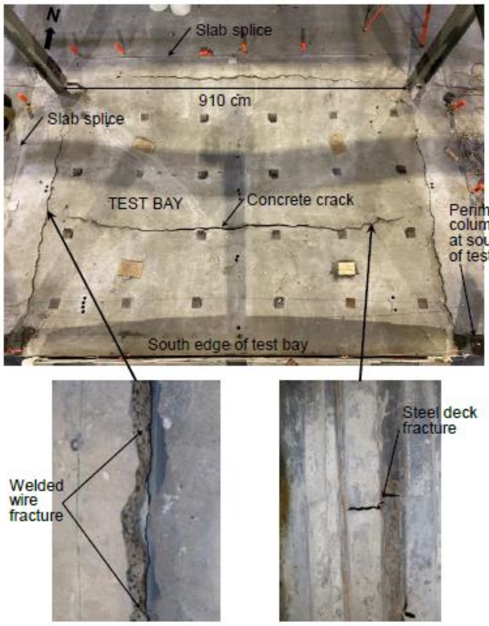


Figure 13 Concrete crack (after cooling) **Source: NIST**

4.5 Discussion on slab reinforcement

Steel reinforcement used in the test floor slab met the minimum shrinkage requirement of 59 mm^2/m specified in the U.S. design provisions [6], approximately equal to 0.075% of the concrete area above steel decking. However, this amount of slab reinforcement was not sufficient to resist tensile forces developed in the slab as the mid-panel deflection increased to 300 mm (approximately L/30) around 70 min of the fire exposure, prior to the required fire resistance rating of 2 hours.

This requirement of slab reinforcement used in a composite floor slab was much smaller than that used in European and New Zealand construction practice or their fire testing (Table 1). The shrinkage steel requirement for reinforced concrete structure [19] is 2.4 times greater than that used in composite floors with steel decking supported by steel floor beams.

In Test 7 [2] of the Cardington test program conducted on a 9 m × 6 m edge bay, similar to the test floor used in the present study, the mid-panel displacement increased to a maximum value of 900 mm (L/10) without collapse. Although the secondary beam remained unprotected and lost much of its flexural capacity during the test, the composite floor with the reinforcement of 142 mm^2/m resisted the imposed load (equivalent to 56% of live loads at ambient temperatures) through tensile membrane action.

Table 1 Shrinkage steel requirements in code provisions and reinforcement used in the present study and the Cardington test.

Reference	Reinforcement ratio (%)	Minimum area/length (mm^2/m)	Spacing (mm)
SDI [6] & Present study	0.075	59	150 typ.
ACI 318 [20]	0.18	165	305 typ.
Eurocode 4 [21]	0.1	80	–
Cardington [2]	0.17	142	200

The Slab Panel Method (SPM)[3], the Steel Construction New Zealand (SCNZ) design software program developed by Clifton et al. [22], was used to predict the behavior of the floor assembly tested in this study. This program incorporates an updated tensile membrane model from that proposed by Bailey [23-24] and the yield line theory by Park [25]. This program predicts the load carrying capacity of a heated slab panel including the secondary beam(s) using the principle of virtual work and also allows for inclusion of the edge reinforcement depending on its ductility. Developed as a design tool, it was used iteratively in this study to generate the results shown.

Figure 14 shows the measured bottom flange temperature of the secondary beam as a function of the mid-panel displacement compared with that predicted using the SPM. The temperature of the secondary beam beyond the fire exposure time longer than 107 min was estimated based on the linear regression of temperature data recorded between 60 min and 107 min. The profiled steel decking was included as a second layer of slab reinforcement (1182 mm^2/m) in addition to welded wire fabric (59 mm^2/m). As shown, the SPM conservatively estimates the mid-panel displacement at secondary beam bottom flange temperature under 930 °C. The SPM prediction implies that although the fire was extinguished in the test

[3] *A certain commercial entity identified herein is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that this entity is necessarily the best available for the purpose of this study.*

at 930 °C, the test floor assembly could resist at higher temperature without collapsing. The predicted maximum temperature at which the test floor assembly failed to resist the applied load was 1025 °C and the corresponding displacement was 627 mm (L/14.5).

Figure 15 shows a further comparison of demand-to-capacity ratio (DCR) as a function of the bottom flange temperature of a secondary beam, estimated using the SPM. The values of DCR greater than 1 imply failure of the floor assembly. The results presented are only applicable to a composite floor assembly similar to the tested specimen. Two different amounts of embedded reinforcement were compared, including 59 mm^2/m with cold-formed plain wires spaced at 150 mm, the same as the welded wire fabric used in the present study, and 230 mm^2/m with No. 3 hot-rolled reinforcing bars (diameter = 9.5 mm) spaced at 305 mm, which could provide larger ductility and satisfies the minimum reinforcement requirement for reinforced concrete (Table 1). The results suggest that a floor assembly with more reinforcement with larger ductility exhibit a superior performance with increasing temperature.

At elevated temperatures, the decking is not typically considered in routine design using the SPM because of the conservative assumption that there is no force transfer between adjacent sheets of decking within the slab panel. However, in this study, the decking was assumed to be continuous across the slab panel since the steel decking exhibited only local failure and maintained its overall integrity based on the post-test inspections. For comparison purposes, hence, the decking was included as reinforcement in the direction that trapezoidal deck ribs ran continuous. However, the influence of steel decking and slab reinforcement on the integrity of a floor system in fire will need to be verified through further studies.

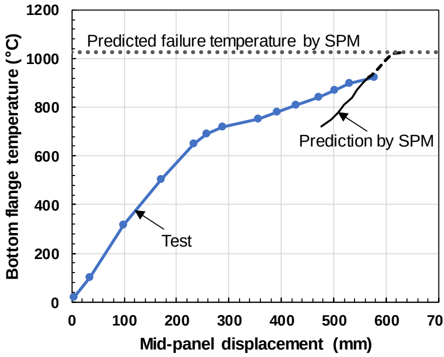


Figure 14 A comparison of the test result with the post-test prediction by SPM

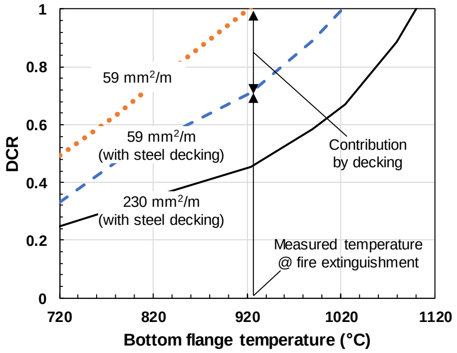


Figure 15 Estimated DCR ratio with respect to bottom flange temperature using SPM

5 Summary and Conclusions

This paper presents a brief overview of the first experiment conducted on a 6.1 m by 9.1 m composite floor assembly of a two-story test building constructed at NIST. Some test results are presented including upper layer gas temperature within the test compartment and responses of the floor assembly to a fire. Also, this paper discusses the influence of slab reinforcement on the behavior and load-bearing capacity of the tested floor assembly.

The experiments showed that the heated floor assembly continuously deflected downward with increasing temperatures. In the early phase of the fire loading, concrete fractured along the perimeter of the test bay. Then, a center longitudinal crack appeared on the floor slab around 70 min due to excessive tension. The mid-panel displacement reached about 580 mm (equivalent to L/16) without collapsing. Fire and mechanical loading were removed around 107 min. The area and ductility of reinforcement used in the test floor slab, the minimum value required in the U.S. construction practice, was not sufficient to delay tensile failure in the slab before 2 hours in a standard fire. According to the post-test prediction using the SCNZ's SPM, the test floor assembly could resist the imposed load slightly longer than 107 min or at higher temperatures due to additional load-carrying mechanism through steel decking although a further study is needed. The use of increased amount of ductile reinforcement (around 0.2-0.3% reinforcement ratio) could significantly improve the fire resistance of a composite floor assembly.

The test data and results partially presented herein can serve as technical information to better understand the behavior of a full-scale composite floor system for improvement of fire safety design provisions and for validation of predictive models used in structural fire engineering.

6 Acknowledgements

The authors would like to thank the support of NIST colleagues on this work: Matthew Hoehler, Matthew Bundy, William Grosshandler, Brian Story, Laurean DeLauter, Anthony Chakalis, Philip Deardorff, Marco Fernandez, Artur Chernovsky, Mina Seif, John Gross, and Xu Dai. The authors would also like to acknowledge the support of expert panel members including William Baker, Glenn Bell, Craig Beyler, Luke Bisby, Ian Burgess, Charlie Carter, Michael Engelhardt, Graeme Flint, Nestor Iwankiw, Kevin LaMalva, Roberto Leon, Jose Torero, and Amit Varma. The authors would like to thank the American Institute of Steel Construction Task Committee 8 (Farid Alfawakhiri and Bob Berhinig) for their independent check on the fire resistance design of the test floor assembly.

References

[1] British Steel. (1999) The behaviour of multi-storey steel framed buildings in fire. United Kingdom, Rotherham.

[2] Wald, F; Silva, L.; Moore, D.; Lennon, T.; Chladná M.; Santiago A.; Beneš, M.; Borges, L. (2006) Experimental behaviour of a steel structure under natural fire. *Fire Safety.* **41**, 509-522.

[3] Bundy, M.; Hamins, A.; Gross, J.; Grosshandler, W.; and Choe, L. (2016) Structural fire experimental capabilities at the NIST National Fire Research Laboratory. *Fire Technology,* **52**, 959-966.

[4] American Society of Civil Engineers (ASCE). (2016) Minimum design loads for buildings and other structures, *ASCE 7*, Reston,

Virginia, USA.

[5] American Institution of Steel Construction (AISC). (2016) Specification for structural steel buildings, *AISC 360*, Chicago, Illinois, USA.

[6] Steel Deck Institute (SDI). Standard for composite steel floor deck - slabs. *C-2011*, USA.

[7] American Society for Testing and Materials (ASTM). (2015) Standard Specification for Structural Steel Shapes, *A992/A992M*, ASTM International, USA.

[8] ASTM. (2019) Standard Specification for Carbon Structural Steel, *A36/A36M*, ASTM International, USA.

[9] ASTM. (2019) Standard Specification for High Strength Structural Bolts and Assemblies, *F3125/F3125M*, ASTM International, USA.

[10] Ramesh, S.; Choe, L.; Seif, M.; Hoehler, M. et al. (2019) Compartment Fire experiments on Long-Span Composite-Beams with Simple Shear Connections: Part1, *NIST Technical Note 2054*, NIST, Gaithersburg, Maryland, USA.

[11] ASTM. (2019) Standard Specification for General Requirements for Steel bars, Carbon, and Alloy, Hot-Wrought, *A29/A29M*, ASTM International, USA.

[12] ASTM. (2019) Standard Test Methods for Fire Tests of Building Construction and Materials, ASTM International, USA.

[13] Zhang, C.; Grosshandler, W; Sauca, A; Choe, L. (2019) Design of an ASTM E119 fire environment in a large compartment. *Fire Technology*, doi: 10.1007/s10694-019-00924-7.

[14] Hamins, A. P.; Maranghides, A.; McGrattan, K. B; Ohlemiller, T.J.; Anleitner, R. (2008) Experiments and Modeling of Multiple Workstations Burning in a Compartment. Federal Building and Fire Safety Investigation of the World Trade Center Disaster, *NIST NCSTAR 1-5E*, NIST, Gaithersburg, Maryland, USA.

[15] Vassart, O. (2014) EN1991-1-2 Basic design methods and worked examples, Chapter 1 in Eurocodes: Background & Applications Structural Fire Design, *Report EUR 26698 EN*, European Union.

[16] Peacock, R.D.; Reneke, P.A.; Forney, G.P. (2017) CFAST-consolidated model of fire growth and smoke transport (version 7), volume 2: user's guide. NIST, Gaithersburg, Maryland, USA.

[17] McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., Weinschenk, C., & Overholt, K. (2013) Fire Dynamics Simulator, User's Guide. NIST, Gaithersburg, Maryland, USA, and VTT Technical Research Centre of Finland, Espoo, Finland.

[18] Sauca, A; Zhang, C; Grosshandler, W; Choe, L; Bundy, M. (2019) Development of a Standard Fire Condition for a Large Compartment Floor Assembly, *NIST Technical Note 2077*, Gaithersburg, Maryland, USA.

[19] International Organization of Standardization (ISO). (1999) Fire-resistance tests — Elements of building construction — Part 1: General requirements, ISO 834, Geneva, Switzerland

[20] American Concrete Institute (ACI). (2019) Building Code Requirements for Structural Concrete. *ACI 318*, Farmington Hills, Michigan, USA.

[21] European Committee for Standardzation (CEN). (2004) Design of Compostie Steel and Concrete Structures Part 1-1: General Rules and Rules for Buidlings. Eurocode 4, European Union.

[22] Clifton, G; Gillies, A.; Mago, N. (2010) The slab panel method: Design of composite floor systems for dependable inelastic response to severe fires, Int. Conference on 6th Structures in Fire: Lancing, Michigan, USA, 492-499.

[23] Bailey, C. (2000) Design of steel members with composite slabs at the fire limit state. Building Research Establishment, United Kingdom.

[24] Bailey, C (2014). Membrane action of slab/beam composite floor systems in fire. *Engineering Structure*, **26**, 1691-1703.

[25] Park, R. (1970) Ultimate Strength of Reinforced Concrete Slabs, Volume 2. The University of Canterbury: Christchurch, New Zealand.

TEMPORAL EXEMPLAR CHANNELS IN HIGH-MULTIPATH ENVIRONMENTS

Mohamed Kashef[⋆] *Peter Vouras*[†] *Robert Jones*[‡] *Richard Candell*[⋆] *Kate A. Remley*[‡]
[⋆] Engineering Laboratory, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, United States
[†] Communications Technology Laboratory, NIST, Gaithersburg, Maryland, United States
[‡] Communications Technology Laboratory, NIST, Boulder, Colorado, United States

ABSTRACT

Industrial wireless plays a crucial role in cyber-physical system (CPS) advances for the future vision of smart manufacturing. However, industrial wireless environments are different from each other and are different from home and office environments. Hence, industrial wireless channel modeling is essential for the development of industrial wireless systems. Moreover, millimeter-wave (mmWave) wireless bands have a high potential to be used for the high data-rates required for industrial automation reliability, with multiple antennas envisioned to mitigate the high path loss. As a result, in this work, we introduce a machine learning (ML) based exemplar extraction approach on mmWave wireless spatial-channel measurements. The proposed approach processes the measured power-angle-delay-profiles to cluster them into a number of groups with respect to the angle of arrival. Then, an exemplar power-delay-profile (PDP) is extracted to represent each group. The resulting set of exemplars provide a tractable way to conduct mmWave industrial wireless systems testing and evaluation by compactly representing various feature groups. This allows the assessment of wireless equipment over the exemplars without the need to test over all of the different instances of wireless channel paths.

Index Terms— Channel modeling, clustering, exemplar channel, industrial wireless, unsupervised learning

1. INTRODUCTION

In future industrial systems, wireless-communication technologies will play a critical role in achieving massive connectivity between various operational components and allowing easier equipment mobility. Industrial physical environments are different than office and home environments which leads to different wireless channel characteristics such as the achievable delay and reliability [1,2]. Generic models are being studied for indoor industrial channels such as [3] where four different categories of wireless channels in indoor factories are considered. However, various industrial environments differ from each other in their layouts, types of equipment, and the performed industrial activities. Hence, designing and testing of industrial wireless systems requires knowledge of the channel characteristics of the corresponding environment [4].

The limited availability of sub-6 GHz wireless spectrum motivated the utilization of millimeter-wave (mmWave) bands for many new wireless technologies. In July 2016, the United States Federal Communications Commission (FCC) allocated 3.85 GHz of licensed spectrum near 28, 37, and 39 GHz for 5G mobile networks, and 7 GHz of unlicensed spectrum from 64–71 GHz that is adjacent to the existing 57–64 GHz unlicensed Industrial, Scientific, and Medical (ISM) bands. The channel characteristics in the mmWave bands are different than sub-6 GHz bands. Moreover, with many licensed bands, they offer a potential candidate for industrial wireless. As a result, channel models are required based on measurements taken in various environments to study the behavior of several technologies and equipment communicating in the mmWave bands.

The power delay profile (PDP) of a wireless channel captures the temporal variations of the channel due to multipath components (MPCs) [5]. The classification and clustering of wireless channels, depending on the extracted temporal features from the corresponding PDPs using machine learning (ML) approaches, have been studied in the literature such as in [6–10]. In these papers, both supervised and unsupervised learning were used for scenario identification. To the best of our knowledge, we are the first to use a clustering ML approach to study the directional impact of the wireless channel and obtain exemplars that can be used to test devices and technologies while representing various groups of the measured directions. The contribution of the paper is introducing the approach, and results that emphasize benefits of applying the innovative approach over realistic channel measurements.

In this work, we introduce an unsupervised learning approach to obtain exemplar PDPs from mmWave channel measurements. Synthetic-aperture measurements are used to determine the channel's power-angle-delay profile (PADP), which characterizes both the angle-of-arrival and time-of-arrival of received power. Synthetic apertures typically consist of a single antenna element that is scanned over multiple locations and the complex fields are acquired over the lattice. The spatial characteristics of the channel are reconstructed in post processing by combining measurements and applying various phasing combinations, much like a phased array. Synthetic apertures have a number of benefits for characterizing spatial channels, including high dynamic range, lower reflectivity (due to the use of a single antenna element), and lower cost.

The proposed approach serves as a way to compactly represent various directional feature groups. Moreover, the result of the proposed approach allows the test and assessment of wireless equipment over the exemplars without the need to test over all of the different instances of wireless channel paths or to evaluate the performance over a generic model that does not capture the specifics of a certain environment. This approach can capture the environmental cases that may stress wireless device performance and potentially reveal flaws or deficiencies that can be improved upon in future device design iterations. To this end, we chose a very challenging propagation scenario inside the Central Utility Plant (CUP) that generates a lot of dense multipath reflections. Unsupervised ML is used to partition the measured PDPs into clusters that correspond to different directions and to extract canonical PDPs that embody all the salient features of each cluster. These exemplars can be used as device stressing design points that can be well replicated in millimeter wave test chambers or within simulation frameworks to test actual device performance. The exemplars generalize the characteristics of measured site-specific data without having to resort to a statistical averaging approach that ultimately yields only benign channel conditions but also requires extraordinary amounts of measured data. Another advantage of the proposed ML approach is that it enables the capability to emphasize or de-emphasize the environmental features that might be more or less important for testing particular device capabilities.

Hany, Mohamed; Vouras, Peter; Jones, Rob; Candell, Rick; Remley, Kate. "Temporal Exemplar Channels in High-Multipath Environments."
Presented at 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2021). June 06, 2021 - June 12, 2021.

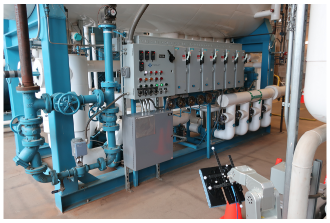


Fig. 1. The CUP measurement environment

2. PROBLEM DESCRIPTION

In this section, an overview of the data collection process is presented. The data preparation and the format of the resulting data are described. Finally, the addressed problem in this work is stated.

2.1. Environment and Data Collection

Measurements were performed in the highly reflective CUP at the Department of Commerce Boulder Laboratories in 2019. This environment consists of large boiler tanks, piping, and numerous racks of control hardware, as shown in Fig. 1. The vector network analyzer (VNA) was placed in a small rack located between the transmit horn antenna and the synthetic aperture receive array. The results presented in this paper were obtained from measurements over the WR-28 waveguide band of 26.5 GHz to 40 GHz with a dynamic range typically on the order of about 90 dB. We positioned a directional horn transmit antenna to point directly at a bank of switches and tanks, as shown in Fig. 2. The horn antenna provides 17 dBi gain, is linearly polarized, and has a 23°/24° 3 dB beamwidth in the E/H planes. It was oriented upward at an elevation angle of approximately 15°. Our synthetic aperture receive array was oriented toward the switch bank as well, minimizing the line-of-sight component. We configured the synthetic aperture [11] to scan a 35-by-35 planar grid with 3mm spacing between the sample points ($\lambda/2$ at 40 GHz). The minimum beamwidth of the synthetic aperture array response is 2.9° attained at 40 GHz in the boresight direction. Since the array is square, the azimuth and elevation beamwidths are equal. With this, we reconstructed directional PDPs corresponding to a beamwidth of 2.9° in azimuth and elevation.

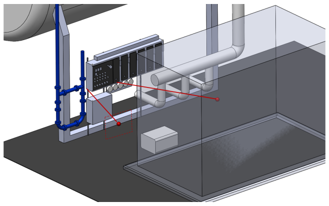


Fig. 2. CAD model of the CUP environment. The red lines show the boresight directions for the transmit antenna and synthetic aperture array. The transparent rectangular prism corresponds to the boiler that is opposite of the control panel. The dotted red rectangle shows the planar synthetic aperture in the environment. The white rectangular prism on the ground corresponds to the VNA.

2.2. Data Preparation and Resulting Data

The S_{21} parameters are collected by the synthetic-aperture channel measurement system in 10 MHz frequency increments between 26.5 and 40 GHz at every spatial sample. The S_{21} parameters are processed using true time delay beamforming to steer the array mainbeam as described in [11]. Initially, a low-sidelobe taper is applied across the aperture that is frequency invariant in the boresight direction. Then to steer the array mainbeam towards a desired direction, an additional phase taper is applied across the aperture that varies linearly with frequency. After coherently combining the product of measured S_{21} values and complex beamforming weights across all the aperture spatial samples, an inverse Fourier transform is utilized to transform the frequency domain data to the temporal domain. The result is known as a directional PDP.

Often, it is desirable that the sum of the power patterns of all the individual beams yields omnidirectional gain. Hence, the pointing directions specified at the peak of the mainbeam are chosen systematically using the approach described in [12] such that all beams overlap at the 3-dB beamwidth. This algorithm accounts for the fact that the width of the array mainbeam increases in proportion to the product of the cosines of the azimuth and elevation scan angles.

2.3. Problem Statement

We denote the PADP instants by $h(\theta, \phi, \tau)$, where θ and ϕ denote the azimuth and elevation of the angle of arrival, respectively, and τ is the delay. The collected data are sampled versions of the PADP, where θ and ϕ take discrete values in the ranges $\theta_{\text{min}} \leq \theta \leq \theta_{\text{max}}$ and $\phi_{\text{min}} \leq \phi \leq \phi_{\text{max}}$. As a result, the input of our problem is a set, $\mathcal{H}$, of the PDPs $h(\theta, \phi, \tau)$ for all fixed combinations of θ and ϕ. The output of the proposed approach is N disjoint groups $\mathcal{H}_i, i \in \{1, N\}$ using unsupervised ML clustering based on a set of defined features. Each group is represented by an exemplar PDP denoted by $\hat{h}_i(\tau)$, which represents the corresponding group $\mathcal{H}_i$.

3. MACHINE LEARNING APPROACH

In this section, the PDP exemplar extraction approach is detailed including the feature selection process and the ML clustering scheme.

3.1. Feature Extraction

Feature-based clustering generally can work with a large set of the PDPs features while it is only performed over a small set of features, in this work, to examine the approach and allow easier visualization of the results. The PDP for a certain pair of θ and ϕ is defined as

$$h(\theta, \phi, \tau) = \sum_{l=0}^{L-1} \alpha_l \delta(\tau - \tau_l),$$
$$\theta_{\text{min}} \leq \theta \leq \theta_{\text{max}}, \phi_{\text{min}} \leq \phi \leq \phi_{\text{max}} \quad (1)$$

where α_l is the power gain for the l-th path, τ_l is the path arrival time, L is the number of the arrival paths, and $\delta(\tau)$ is the Dirac function.

Now, we define the features of $h(\theta, \phi, \tau)$ that we used to characterize all the directional PDPs. All of the features are evaluated for a single PDP at a certain pair of θ and ϕ. We will drop the argument to simplify the expressions. The channel gain, G, is evaluated as

$$G = \sum_{l=0}^{L-1} \alpha_l. \quad (2)$$

The dynamic range of the PDP is defined as the ratio of the largest MPC to the smallest MPC. It is evaluated in dB as

$$R = 10 \log \frac{\max_{0 \leq l \leq L-1}(\alpha_l)}{\min_{0 \leq l \leq L-1}(\alpha_l)}. \quad (3)$$

Hany, Mohamed; Vouras, Peter; Jones, Rob; Candell, Rick; Remley, Kate. "Temporal Exemplar Channels in High-Multipath Environments."
Presented at 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2021). June 06, 2021 - June 12, 2021.

The mean delay is the first moment of the power delay profile and is evaluated as follows

$$\tau_{\text{mean}} = \frac{1}{G}\sum_{l=0}^{L-1} \alpha_l \tau_l. \tag{4}$$

The root-mean-squared (RMS) delay spread is the second moment of the power delay profile and is evaluated as follows

$$\tau_{\text{RMS}} = \sqrt{\frac{1}{G}\left(\sum_{l=0}^{L-1} \alpha_l \tau_l^2 - \tau_{\text{mean}}^2\right)}. \tag{5}$$

3.2. Data Clustering

The first step in clustering is to normalize the features to the range of 0 to 1. Then, we perform spectral clustering over the vectors of features corresponding to each beam direction. Generally, spectral clustering is a powerful technique in clustering nonlinear and inseparable data and where non-convex clusters are allowed [13]. We have used the Scikit-learn implementation for the spectral clustering algorithm [14]. We denote the feature vector, corresponding to a PDP in a specific beam direction defined by a pair of azimuth and elevation angles, by Z_j, where j is the index of the PDP. The length of the vector Z_j is the number of features, N_{f}, and the total number of PDPs to be clustered is N_{H}.

The spectral clustering algorithm requires a similarity metric and a number of output clusters. In this work, we consider linear similarity although other similarity metrics can be considered as well. The pairwise linear similarity between feature vectors evaluated through their dot product may be given as

$$SL_{jk} = Z_j^{\text{T}} Z_k, \tag{6}$$

where $(*)^{\text{T}}$ is the transpose of the vector.

Finally, the number of the clusters is obtained through a recursive search for the maximum Silhouette score [15]. The Silhouette score measures how closely-related an object is to its own cluster against the other clusters. We repeat clustering over various values of N and keep the clusters that achieve the highest Silhouette score.

3.3. Exemplar Extraction

The last phase of the approach is to extract an exemplar PDP from each of the clusters. The exemplar is a member of a cluster of PDPs. The initial used scheme defines the exemplar as the PDP with shortest mean feature-distance to other cluster members. In a cluster $\mathcal{H}_i$, the exemplar PDP, denoted by $\hat{h}_i(\tau)$, is the one corresponding to the feature vector $\hat{Z}_i$. The value of $\hat{Z}_i$ is evaluated as

$$\hat{Z}_i = \underset{Z_k \text{ s.t. } k \in \mathcal{H}_i}{\arg\min} \frac{1}{|\mathcal{H}_i| - 1} \sum_{\substack{Z_j \neq Z_k \\ Z_j \text{ s.t. } j \in \mathcal{H}_i}} \sqrt{||Z_k - Z_j||^2}. \tag{7}$$

This exemplar represents the closest cluster member to the center of mass of the cluster where the masses of all the clusters members are equal.

Moreover, various weighting factors can impact the exemplar extraction process, where the mass of each member of the cluster can vary based on the importance of various features. One example is to scale the cluster members' mass based on their total channel power gain value, G. Hence, we introduce, to the exemplar extraction process, a channel-gain based scaling with a power-exponent tuning. We define power-exponent σ to be the exponent to the PDP power gain G that is multiplied by the minimization argument in (7). The extracted exemplar in this case with a power exponent of σ is evaluated as follows

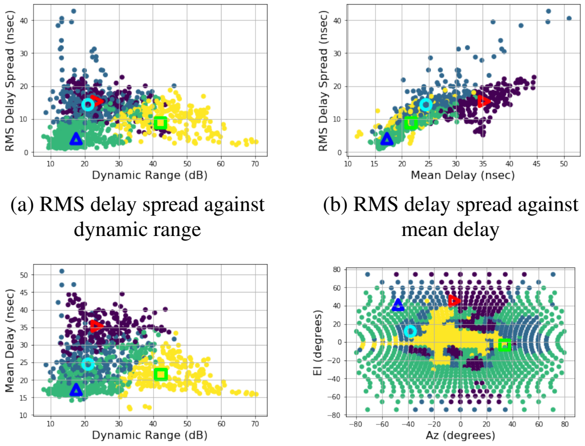
$$\hat{Z}_i = \underset{Z_k \text{ s.t. } k \in \mathcal{H}_i}{\arg\min} \frac{G_k^{\sigma}}{|\mathcal{H}_i| - 1} \sum_{\substack{Z_j \neq Z_k \\ Z_j \text{ s.t. } j \in \mathcal{H}_i}} \sqrt{||Z_k - Z_j||^2}, \tag{8}$$

where G_k is the total channel-power gain of the PDP corresponding to Z_k. For high power-exponent values, this scheme will pick the exemplar with the highest channel-power-gain value in the cluster.

4. RESULTS

In this section, we show the results obtained by running the exemplar extraction approach over the channel measurements. Two sets of results are included: i) Linear similarity metric with $\sigma = 0$, and ii) Linear similarity metric with $\sigma = 0.5$. In each set of results, we show two figures. The first shows the features of the resulting clusters through four scatter plot sub-figures: a) RMS delay spread (ns) against dynamic range (dB), b) RMS delay spread (ns) against mean delay (ns), c) mean delay (ns) against dynamic range (dB), and d) the directions in azimuth and elevation (degrees) of clusters. All cluster members have the same color with a corresponding exemplar marked with a specific shape. In the second figure, the PDPs of the exemplars are shown.

4.1. Linear Similarity with $\sigma = 0$



(a) RMS delay spread against dynamic range

(b) RMS delay spread against mean delay

(c) Mean delay against dynamic range

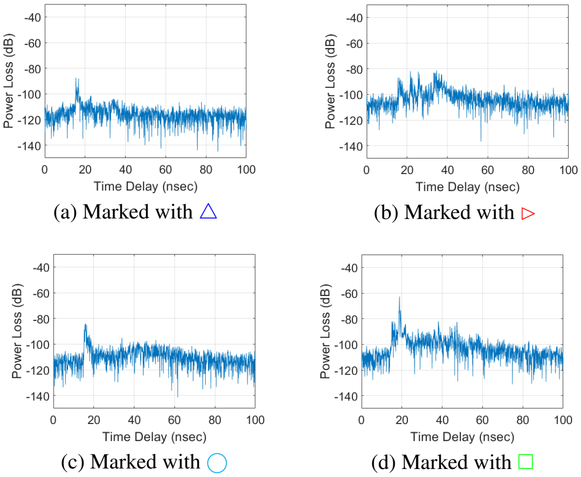
(d) Azimuth against elevation angles

Fig. 3. Clustered data with marked exemplars at $\sigma = 0$

In Fig. 3, we draw the clustering result with the exemplars marked on the figures. In this case, we found that the optimal number of clusters is four, where the Silhouette score is 0.32. The value of Silhouette score can be used to demonstrate the effectiveness of the proposed clustering approach, because it measures how closely related a directional PDP is to its own cluster against the other clusters with respect to the used features.

The first cluster, in green, is characterized by low mean delay, low RMS delay spread, and low dynamic range. The second cluster, in purple, is characterized by high mean delay, low RMS delay spread, and low dynamic range. The third cluster, in blue, is characterized by high RMS delay spread, and low dynamic range.

The fourth cluster, in yellow, is characterized by low mean delay, low RMS delay spread, and high dynamic range. The corresponding exemplars are marked in Fig. 3 with different shapes, namely, up-pointing triangle, right-pointing triangle, circle, and square, respectively. The PDPs of these exemplars are shown in Fig. 4. This figure shows that an exemplar PDP reflects the characteristics of the corresponding cluster. As an example, in Fig. 4(a), the PDP reflects the case where the PDP has a single peak, which leads to low mean delay and low RMS delay spread. This peak also has a low channel gain such that the dynamic range is low as well.



(a) Marked with △

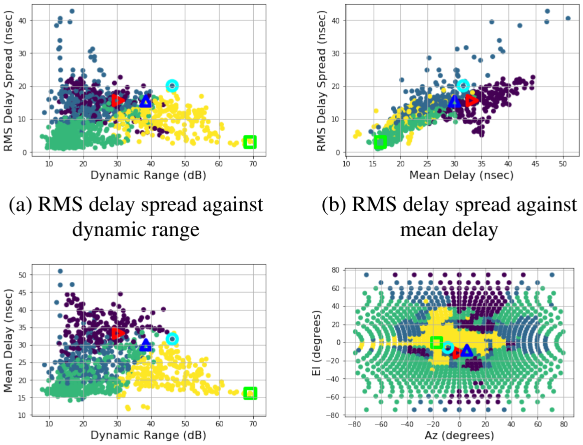
(b) Marked with ▷

(c) Marked with ○

(d) Marked with □

Fig. 4. Marked exemplar PDPs corresponding to Fig. 3

4.2. Linear Similarity with $\sigma = 0.5$



(a) RMS delay spread against dynamic range

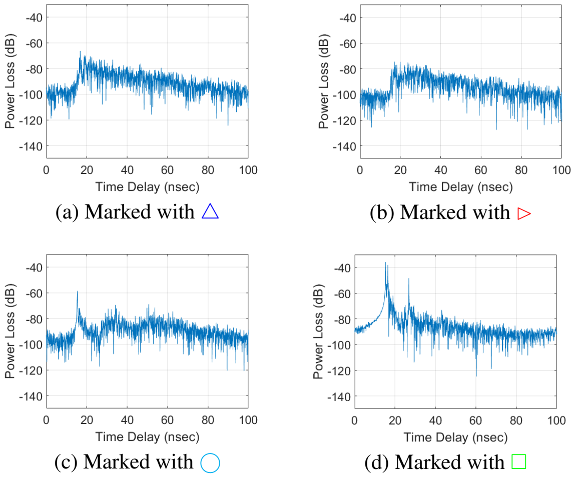
(b) RMS delay spread against mean delay

(c) Mean delay against dynamic range

(d) Azimuth against elevation angles

Fig. 5. Clustered data with marked exemplars at $\sigma = 0.5$

In this set of results, we present in Fig. 5 similar clusters as those obtained in Fig. 3. However, the increase in the exemplar extraction power exponent allows us to extract exemplars with higher channel-power gain and, hence, higher dynamic range. The clusters for the case of $\sigma = 0.5$ are shown in Fig. 5 and the corresponding exemplars are shown in Fig. 6, where they exhibit higher gain in the MPCs than the MPCs in Fig. 4. Note that the exemplars in Fig. 6 (a) and (b) look similar because the change in power-exponent moved the exemplars to the edges of the clusters and hence close to each other.



(a) Marked with △

(b) Marked with ▷

(c) Marked with ○

(d) Marked with □

Fig. 6. Marked exemplar PDPs corresponding to Fig. 5

4.3. Comparison to the Average PDP

In Fig. 7, we show the average PDP of all the measured PDPs as a benchmark for representing the measured data. Comparing the exemplars to the average PDP, we observe that each exemplar represents a group of PDPs with certain features and hence can be used for testing wireless devices under certain conditions.

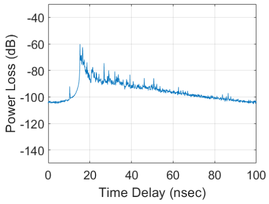


Fig. 7. The average PDP over all the directional PDPs

5. CONCLUSIONS

In this paper, we introduced an approach for directional PDP exemplar extraction from measured data for a static, highly reflective channel. The approach deploys unsupervised spectral clustering for PDP clustering and uses the delay and power characteristics for exemplar extraction. We have demonstrated that the wireless channel paths between two points in an industrial environment can have different characteristics depending on orientation. Hence, to operate in such an environment, a wireless node has to be tested under all different types of channel characteristics. The proposed approach serves as a way to compactly represent various feature groups. This allows the test and assessment of wireless equipment over the exemplars without the need to test over all of the different instances of wireless channel paths or to evaluate the performance over a generic model that does not capture the specifics of a certain environment.

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.

6. REFERENCES

[1] Kay Soon Low, Nu Win, and Meng Joo Er, "Wireless Sensor Networks for Industrial Environments," 2005.

[2] M. Cheffena, "Propagation channel characteristics of industrial wireless sensor networks [wireless corner]," *IEEE Antennas and Propagation Magazine*, vol. 58, no. 1, pp. 66–73, Feb. 2016.

[3] 3GPP, "Addition of indoor industrial channel model to Doc. 38.901 - Study on channel model for frequencies from 0.5 to 100 GHz," Tech. Rep. R1-1909807, 3rd Generation Partnership Project (3GPP), 09 2019.

[4] R. Candell, M. Kashef, Y. Liu, K. B. Lee, and S. Foufou, "Industrial wireless systems guidelines: Practical considerations and deployment life cycle," *IEEE Industrial Electronics Magazine*, vol. 12, no. 4, pp. 6–17, Dec. 2018.

[5] B. Soret, M. C. Aguayo-torres, and J. T. Entrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated Rayleigh channel," *IEEE Transactions on Wireless Communications*, vol. 9, no. 6, pp. 1901–1911, June 2010.

[6] J. Zhang, L. Liu, Y. Fan, L. Zhuang, T. Zhou, and Z. Piao, "Wireless channel propagation scenarios identification: A perspective of machine learning," *IEEE Access*, vol. 8, pp. 47797–47806, 2020.

[7] M. I. AlHajri, N. T. Ali, and R. M. Shubair, "Classification of indoor environments for iot applications: A machine learning approach," *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 12, pp. 2164–2168, 2018.

[8] C. Huang, A. F. Molisch, R. Wang, P. Tang, R. He, and Z. Zhong, "Angular information-based NLOS/LOS identification for vehicle to vehicle MIMO system," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[9] Haihan Li, Yunzhou Li, Shidong Zhou, and Jing Wang, "Wireless channel feature extraction via GMM and CNN in the tomographic channel model," *Journal of Communications and Information Networks*, vol. 2, no. 1, pp. 41–51, Mar. 2017.

[10] Yunlong Yu, Fuxian Liu, and Sheng Mao, "Fingerprint extraction and classification of wireless channels based on deep convolutional neural networks," *Neural Processing Letters*, vol. 48, no. 3, pp. 1767–1775, Feb. 2018.

[11] P. Vouras, B. Jamroz, J.T. Quimby, A. Weiss, D.F. Williams, R. Leonhardt, D. Guven, R.D. Jones, J. Kast, and K.A. Remley, "Wideband synthetic aperture millimeter-wave spatial channel measurements with uncertainties," *in review*, 2020.

[12] G. V. Trunk W. M. Waters, D. P. Patel, "Optimum number of faces of a volume-scanning active array radar," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 3, July 1998.

[13] Soheila Ashkezari Toussi and Hadi Sadoghi Yazdi, "Feature selection in spectral clustering," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, no. 3, pp. 179–194, Sep 2011.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] "Online, sklearn.metrics.silhouette_score, URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html," (visited: 2020-09-23).

Apply Quantum Search to the Safety Check for Mono Operational Attribute Based Protection Systems

Vincent C. Hu

Computer Security Division
National Institute of Standards and Technology
Gaithersburg, U.S.A. 20899
vhu@nist.gov

Abstract. Interrelated computing device's system such as IoT, RFID, or edge device's systems are pervasively equipped for today's information application and service systems, protecting them from unauthorized access i.e. safety is critical, because a breach from the device may cause cascading effects resulting to data lost or even crash of the whole information system. However, to determine a protection system's safety is proven to be undecidable unless the system has limited management capabilities. And even with such limitation, it is too expensive to perform a safety test in term of computation time when a device has more than hundreds of subjects which is not uncommon for interrelated computing devices. Nevertheless, the required exponential computing time for safety test can be significantly reduced to its square root if computed by quantum algorithm. In this paper we demonstrate an application of quantum search algorithm to reduce the computation time for safety test for limited (i.e. mono operational) protection systems which are based on attribute-based access control model. The improvement of the performance allows the safety test for interrelated computing device's system to be much less expensive to compute.

Keywords: Access Control; Protection System; Quantum Algorithm; Security Model; Quantum Search.

1 Introduction

Interrelated computing device's systems (ICDSs) [1] such as IoT, RFID, edge device's or similar systems are pervasively equipped for today's information applications and services [2], protecting these systems from unauthorized access i.e. safety is critical, because a breach from an ICDS may cause cascading effects leading to data lost or even crash of the whole system[3]. Further, as Attributed-Based Access Control (ABAC) [4] model is getting more applied for access control [5], an ICDS may apply ABAC for its access control mechanism. We call such protecting system the **Attribute Based Protection System** (ABPS), which also includes its access control **policy management functions**.

Hu, Vincent C. "Apply Quantum Search to the Safety Check for Mono Operational Attribute Based Protection Systems." Presented at 13th International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage (SPACCS 2020), Nanjing, CN. October 23, 2020 - October 25, 2020.

The **safety** for an ABPS is to ensure that it is impossible to leak access **privilege** (perform actions to objects) from authorized subjects to unauthorized subjects through any changes of access state. And **safety test** is to verify if the safety of the system is maintained after any order of access control policy changes. To test that in worst case obviously requires checking access control policy updates evoked by all possible sequences of policy management functions. By HRU[1][6] theory, such test is proven to be undecidable unless the ABPS is limited to be **mono operational**, which is restricted to have only one primitive commend for each policy management function.

An ICDS's ABPS can be mono operational, because its access control requires limited management capability. But even that, the exponential computation time (NP-Complete) for safety test is still too expensive [7], because subjects, objects, and actions as exponential variables of computing time for most ICDSs can easily reach to hundreds if not thousands. Nevertheless, it can be significantly improved by applying quantum search algorithm, which reduces the computation time to the square root of the time required by classical algorithm, thus, allows the safety test for ICDS's ABPS to be minimum computable.

This paper is divided into six sections, section I is the introduction, Section II describes the ABPS, Section III explains the safety test algorithm that applied to mono operational ABPS, Section IV introduces quantum algorithm modified from quantum search algorithm for the privilege leak detect process of safety test algorithm. Section V demonstrates the performance comparison between quantum and classical algorithms in terms of computation time, and Section VI is the conclusion.

2 Attribute Based Protection System

Attribute Based Access Control (ABAC) is an access control method where subject requests to perform actions on objects are granted or denied based on assigned attributes of the subject and objects, environment conditions, and a set of rules specified by those attributes and conditions [4] called ABAC **policy**, which given the values of the attributes of the subject, object, and environment conditions and their relations make it possible to determine if a requested access should be authorized.

ABPS applies ABAC where a **subject** s_i represents a combination of **subject attributes** $sa_1,\ldots sa_i,\ldots sa_k$ the subject is associated with, and an **object** o_i represents a combination of **object attributes** $oa_1,\ldots oa_j,\ldots oa_l$ that apply to the object. And the access control policy is managed by the policy management functions. The ABPS's access state can be presented by the **HRU access matrix** (Figure 1) such that the access control policy rules are mapped to rows and columns with intersected cells. A cell contains actions that are permitted to perform the accesses from the subject to the object corresponding to the row and the column, as example in Figure 1, shows that subject s_j is permitted to perform actions r and w accesses to object o_i, The cell intersected by both row and column of subjects is used for creating or deleting a subject by another subject. ABPS's ABAC policy rules

[1] We denote the term HRU to be general references to the systems and theories presented in [5].

are mapped to the access matrix by adding permitted access actions into cells and removing denied accesses actions from cells if the actions existed.

Subject\Object	s_1	…	s_n	o_1= (….)	…	o_i = (oa_1,…oa_j,…oa_l)	…	o_m= (….)
s_1 = (….)								
…..								
s_j = (sa_1,..sa_i,..sa_k)						*r, w*		
…..								
s_n = (….)								

Fig. 1. ABPS access matrix state.

An ABPS's policy management mechanism, which in general is a set of policy management functions for creating, updating the ABAC policy rules. The function is intrinsically equal to access **matrix update function** such that *assign rule* function: "*assign action a to object o_j to subject s_k*" is equal to: "*add a to the intersect cell of row s_k and column o_j*" add function, and *delete rule* function: "*delete action a to object o_j of subject s_k*" is equal to: "*remove a in the intersect cell of the row s_k and column o_j*" delete function. Figure 2 shows an example ABAC rule: "*users with attribute p or q can read device x*" maps to HRU access matrix. Therefore, an ABPS access state is an instance of an HRU access matrix state, and ABAC policy rules can be configured to rows and columns of an HRU access matrix.

Users/Device	…..	device *x*	
…….			
Attribute *p*		*read*	
…….			
Attribute *q*		*read*	

Fig. 2. ABPS access matrix state

There are six **primitive commands** for ABPS's policy management functions, and their counter parts for access matrix operations are shown in Table 1:

Table 1. ABPS'S and HRU primitive commands mapping.

ABPS primitive commands	HRU primitive commands
assign action (a, s_i = (sa_1…..sa_k), o_i = (oa_1…..oa_j))	*enter action a into* (s_i, o_i)
delete action (a, s_i = (sa_1…..sa_k), o_i = (oa_1…..oa_k))	*delete action a into* (s_i, o_i)
add subject (s_i = (sa_1…..sa_k))	*create subject* s_i
add object (o_i = (oa_1…..oa_k))	*create object* o_i
remove subject (s_i = (sa_1…..sa_k))	*destroy subject* s_i
remove object (o_i = (oa_1…..oa_k))	*destroy object* o_i

Access state changes after executing sequences of access state change functions can be presented formally by: $Q_1 \vdash fn_1\ Q_2 \vdash$ ….. $fn_i\ Q_i \vdash$ ….. $fn_m\ Q_m$, ($\vdash$ means complete the function) where Q_i is an access state, and function fn_i makes access state change from Q_i to Q_i+1. The pseudo fn_i for ABPS policy management function is:

```
ABPS_fn_i (subjects, actions, objects){//* subjects or objects are optional if the
functions are create/destroy subjects or objects *//
    if no conflict with current ABAC access control policy
    then { execute primitive commands $pc_1$;
    execute primitive commands $pc_2$;
    ….
    execute primitive commands $pc_n$,
    that apply to the subjects, actions, and objects
    }//*primitive commands update ABAC policy *//
    current access control policy $P_i = P_{i+1}$
}
```

And the corresponding fn_i for HRU access matrix update function is:

```
HRU_fn_i (subjects, actions, objects) { //* subjects or objects are optional if the
functions are create/destroy subjects or objects *//
    if conditions $c_1$, $c_2$….$c_k$ then {
    execute primitive commands $pc_1$;,
    execute primitive commands $pc_2$;,
    ….
    execute primitive commands $pc_n$,
    that apply to the subjects, actions, and objects
    } //*primitive commands update access matrix *//
    curent access matrix $H_i = H_{i+1}$
}
```

The steps for checking the conflict of access control policy in *ABPS_fn*$_i$ are not semantically different from HRU_*fn*$_i$'s *if* condition checks, because satisfying the current ABPS policy is the same as satisfying the state of HRU matrix, which can be translated from the ABPS's policy rules.

3 ABPS safety check

HRU defines that:

"*given a protection system, we say command **c** leaks generic action **a** from the access state if **c**, when run on the access state, can execute a primitive operation which enter **a** into a cell of access matrix which did not previously contain **a***".
from the definition, the **safety** of ABPS is to ensure unintended subjects cannot perform protected actions on objects through executing any sequence of policy management functions (*ABPS_fn*$_i$s described in Section II), such that the permitted accesses for the action by the original access control policy remains the same

during the system's life cycle. Thus, the **safety test** is to verify that if the system remains safe after all possible sequence of policy management functions being executed. Figure 3 shows the components and relations of an ABPS and its safety test system.

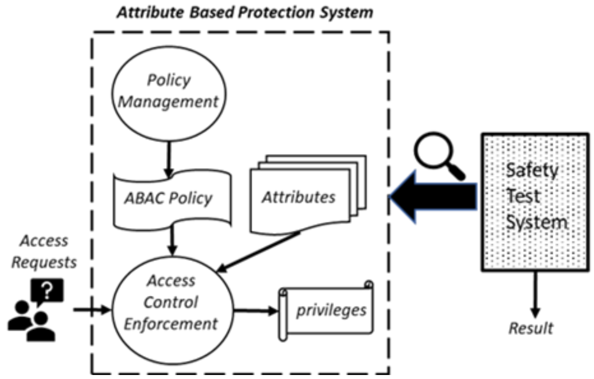


Fig. 3. ABPS and Safety test system

According to HRU, safety test for general protection systems including ABPSs is undecidable in term of computation complexity, because to test the safety of a protection system in worst case obviously requires checking all possible sequences of access matrix changes evoked by access matrix update functions with all possible parameters including subjects, actions, and objects, plus that the function may contain unlimited *if* condition checks and arbitrary numbers of primitive commands (Table 1). The undecidability can be proved by configuring the protection system to simulate the behavior of an arbitrary Turing machine, with a safety leakage state corresponding to the Turing machine entering a final state [6].

In addition to general protection systems, from HRU's theory, a restricted type of protection system called **mono operational protection system**, which limits each matrix update function to contain only one primitive command. HRU shows that determining safety for mono operational protection system is decidable in NP-Complete, which is proved by reducing a K-clique problem to a safety decision problem that translating the system's initial access matrix to an **adjacency matrix** for a graph, then test to see if it forms a *k*-clique before entering an action *a* to the access matrix causing safety leak. HRU also shows that only the primitive command *enter* can change the access state. To simulates the HRU's algorithm for mono operational ABPS safety test, Figure 4 illustrates the algorithm *Safety_Test* for an action *a*. Since a policy management function of a mono operational protection system contains only one primitive command, and only the *enter* primitive command can change the access state, the algorithm needs to tests every possible sequences of *enter* commands for all actions, in other words, try all possible sequences of primitive *enter* commands, (optional starting with a *create* subject command) of length up the $|A|\times|S|\times|O|$ for each sequence, where $|A|$ is the number of all actions, $|S|$ is the number of all subjects, and $|O|$ is the number of all objects.

The parameters of an *enter* command are an action-subject-object triplet corresponding to a command sequence, which is identified by a binary number, for example, if there are two subjects s_1 and s_2, two objects o_1 and o_2 , and two actions a_1 and a_2 in the ABPS then there are 2×2×2 = 8 different *enter* command, and 2^8 possible sequences, for instance, the 5th *enter* command sequence is {*enter* (a_1, s_1, o_1); *enter* (a_1, s_2, o_1)}, and the 24th command sequence is {*enter* (a_1, s_1, o_1); *enter* (a_1, s_2, o_2); *enter* (a_2, s_1, o_1)}, because the binary form of the sequence 5 is 00000101 and 24 is 00011001, where the bits representations of *enter* commands are assigned in Table 2.

Table 2. Example bit number assignment of 2 actions, 2 subjects and 2 objects pairs.

Triplet	*Assigned bit*
(a_1, s_1, o_1)	1st bit
(a_1, s_1, o_2)	2nd bit
(a_1, s_2, o_1)	3rd bit
(a_1, s_2, o_2)	4th bit
(a_2, s_1, o_1)	5th bit
(a_2, s_1, o_2)	6th bit
(a_2, s_2, o_1)	7th bit
(a_2, s_2, o_2)	8th bit

```
Safety_Test ($P_1$, $a$ ) {
    $H_1$ = Initial $P_1$; //* map accesses permitted by the ABPS access control policy to the HRU access matrix *//
    Privilege_leak = 0 ;
    $i$ = 1;
    For $k$ = 0 to $2^{|G|\times|S|\times|O|}$ - 1  //* $|G|$ is the number of actions, $|S|$ is the number of subjects, $|O|$ is the number of objects
    *//{
        For all (a$_x$, $s_i$, $o_j$)  //*( $a_x$, $s_i$, $o_j$) $\in$ {$A\times S \times O$}; $A$ is the set of actions, $S$ is the set of subjects, $O$ is the set of objects *//
        {
            If Bitmap ($a_x$, $s_i$, $o_j$, $k$) = 1  //*match $s_i$-$o_j$ pair to binary number k*//
                enter ($a_x$, $s_i$, $o_j$)
      {
        }
        $H_i$ = $H_{i+1}$;
        If State_Compare ($H_i$, $H_1$, $a$)  //*check if access state is changed*//
        Then {
        privilege_leak = 1;
        end Safety_Test;
        }
        else privilege_leak = 0;
        }
}
Bitmap ($a_x$, $s_i$, $o_j$, $k$) {
        $i$ = Numer_map($a_x$, $s_i$, $o_j$) //* translate $s_i$-$o_j$ pair to binary number*//
        For $j$ = 1 to $i$
        If the $j$th bit of Binary($k$) == 1 //*check the match of bits*//
                return 1
}
State_Compare ($H_i$, $H_1$, $a$){
      For each row of $s_i$ {
            For each column of $o_i$ {
                  If ((($a$ in the cell of ($s_i$, $o_i$)) of $H_i$) == (($a$ in the cell of ($s_i$, $o_i$)) of $H_1$))
                        Then return leak = 0
                        else return leak = 1; //*privilege_leak state is passed to Safety_Test*//
            }
      }
}
```

Fig. 4. ABPS Safety Test algorithm

The *Bitmap* function translates the action-subject-object triplet of the *enter* command to a binary number to match the current sequence number passed to the function as examples showed in Table 2.

The *State_Compare* function compares cells in original access matrix H_1 to the new access matrix H_i that might be updated after a sequence of *enter* commands were executed, it checks if a privilege leak by action a is found, and the result is returned to the *Safety_Test*. Note that the algorithm only checks the safety against one action a, it is capable of checking multiple actions leaks, and to do that we need to replace (s_i, o_j) with (s_i, a_m, o_j), $\{S\times O\}$ with $\{S\times A\times O\}$, $2^{|S|\times|O|}$ with $2^{|S|\times|A|\times|O|}$, and ($a$, s_i, o_j) with (a_m, s_i, o_j) and add a For loop for each a_m check in the function.

For later discussion of quantum algorithm, we call the *For* loop from line 5 to 14 in Figure 4 the *Leak_Detect* process collectively. Hence, the *Safety_Test* would require $2^{|A|\times||S|\times|O|}\times$ O(*Leak_Detect*) computation time (steps) for detecting an access privilege leak, where O(*Leak_Detect*) is the time needed for *Leak_Detect* process, which is equal to O(*Bitmap*)$\times$ $|A|\times|S|\times|O|$+O(*State_Compare*) = $2\times|A|\times|S|\times|O|$, because O(*Bitmap*) take constant and O(*State_Compare*) takes number of steps equals to the size of access matrix: $|S|\times|O|$ times $|A|$ to compute.

Some low power ICDSs' (e.g. IoT, RFID, or edge computing devices or similar systems) access control are managed by ABPS, where access control policies are either embed or deployed by central management system rather than managed by the device themselves [8]. For instance, RFID devices include independent storage access control rules, only when the rule needs to be updated, do reading devices need to communicate with the server, and access control rules can be updated by the multicast method. In the same security zone, multiple reading devices can distribute access control rules at the same time, thereby improving the efficiency of rule updates [9]. In addition, some access control mechanisms allow smart objects take the authorization decisions based on current context of the processes in use [10]. For those systems with limited access control management capabilities, the protection systems can be implemented by mono operational ABPS. And these ICDSs usually accessed by a large number of users risking safety leak [11], plus, due to frequently adding new and updating old devices, their safety need to be efficiently verified to satisfy their security and performance requirements of services, thus, need an efficient safety test method that classical algorithms cannot offer.

4 Quantum search algorithm for ABPS safety check

Even the ABPS safety test is decidable but in NP-Complete as described in the last section, it is still an issue to be efficiently computable for systems having large number of subjects, objects and actions such might sum up to hundreds if not thousands of users, because, for example, an ICDS is used by just 10 subjects (classified by users' attributes) with only two objects (classified by devices' attributes) and 3 actions, the safety will take $2^{10\times2\times3} \times (2\times10\times2\times3)$ computation time. Thus, it is desirable to improve the exponential computation time (steps) to be feasible to compute. To reduce the computation time, we propose to adopt the Grover Quantum search algorithm [12, 13], which performs the transformation $L|x\rangle|q\rangle = |x\rangle|q\otimes f(x) \rangle$ to a black box oracle f to speed up $f(x)$ for multiple x inputs, where $|q\rangle$ is an ancilla qubit for quantum unitary computation. The algorithm finds with high probability the unique input to the black box oracle function that produces a particular output value, using just $\sqrt{N}$ evaluations of the function, where N is the size of the function's domain.

Schema in Figure 5 shows the application of quantum search algorithm for safety test called ***Safety_Test* quantum algorithm**, which uses n+1 qubit register as input (the ancilla 1 qubit is for quantum unitary operation), where $N = 2^n = 2^{|A|\times|S|\times|O|}$ is the number of all possible sequences of *enter* commands, $|A|$ is the number of actions, $|S|$ is the number of subjects, and $|O|$ is the number of objects of the ABPS. The output of the algorithm is a number x_{leak} representing a sequence of *enter* commands that causes privilege leak by the action a. Notice that instead of a leak command sequence, the classical *Safe_Test* algorithm (Figure 4) only returns a result indicating whether a leak exist. In contrast, the quantum algorithm returns one of the leak sequence numbers (there could be more than one command sequence that cause leakages). The black box oracle function f is hence the *Leak_Detect* process (from line 5 to 14 of the classical *Safety_Test* algorithm in Figure 4).

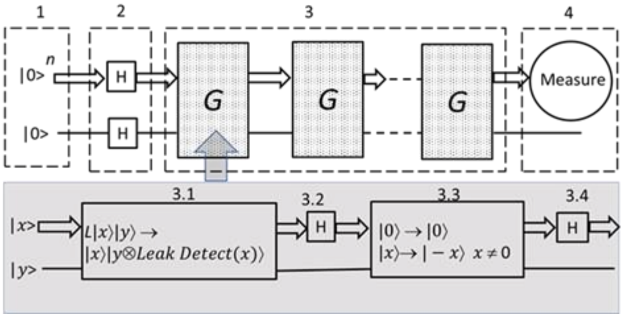

1
2
3
4
|0>
n
|0>
H
H
G
G
G
Measure
3.1
3.2
3.3
3.4
|x>
|y>
$L|x\rangle|y\rangle \rightarrow$
$|x\rangle|y \otimes Leak\ Detect(x)\rangle$
H
$|0\rangle \rightarrow |0\rangle$
$|x\rangle \rightarrow |-x\rangle \ x \neq 0$
H

Fig. 5. Quantum Safety Check schema

Safety_Test quantum algorithm requires repeating applications of the Grover quantum search subroutine shown as the **Grover iteration** *G* in Figure 5, where each iteration move $1/\sqrt{N}$ amplitude towards solutions, thus $\sqrt{N}$ iterations should suffice to render a x_{leak}. The algorithm is divided into four steps as below

1) Begins with the initial state, $n + 1$ qubits in the state $|0\rangle{:}|0\rangle^{\otimes n}|0\rangle$, the extra $|0\rangle$ is for the quantum unitary operation.

2) The Hadamard transform is applied to establish equal superposition state $|\Psi\rangle$ of all possible numbers of *enter* command sequences that

$$|\Psi\rangle = \frac{1}{\sqrt{2^n}} \sum_{x=0}^{2^n-1} |x\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right]$$

Where $0 \le x \le 2^{|A|\times|S|\times|O}$.

3) Apply the Grover *G* iteration $K = \lceil \sqrt{N/M} \rceil$ times: where *M* is the number of sequences of *enter* command sequences (i.e. x_{leak}s) that cause privilege leaks. This step can be subdivided into the following three steps:

3.1) Apply the quantum oracle *L*

$$L|x\rangle|y\rangle = |x\rangle|y \otimes Leak\ Detect(x)\rangle$$

resulting

$$|x\rangle \rightarrow (-1)^{Leak\ Detect(x)}|x\rangle$$

Note that each *x* is a number representing an enter commands sequence, for example the number 5 represent the sequence; { *enter* (g_2, s_1, o_2); *enter* (g_1, s_2, o_1)} as shown in Section III. *Leak_Detect* (*x*) = 0 for all $0 \le x \le 2^n$ except the x_{leak} for which *Leak_Detect*(*x*) = 1 indicating the *enter* command sequence leaks privilege for action *a* in the current access control state.

3.2) Apply the Hadamard transform $H^{\otimes n}$

$$|\Psi\rangle = \frac{1}{\sqrt{2^n}} \sum_{x=0}^{2^n-1} (-1)^{Leak\ Dectect(x)}|x\rangle \left[\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right]$$

3.3) Performs a conditional phase shift i.e.

$$|0\rangle \rightarrow |0\rangle \text{ and}$$
$$|x\rangle \rightarrow |-x\rangle,\ \ x \neq 0$$

with every computational basis state except $|0\rangle$, receiving a phase shift of -1, i.e. the leaking *enter* command sequence $x = x_{leak} \neq 0$. The conditional phase shift can be calculated by applying the matrix operation of

$$2\,|0\,\rangle\langle 0| - I$$

where *I* is the identity matrix.

3.4) Apply the Hadamard transform $H^{\otimes n}$

4) Measure the first *n* qubits of $|\phi\rangle$ gets one of the possible leak sequence x_{leak}.

The quantum algorithm requires $\sqrt{N/M}$ × O(*Leak_Detect*) [14] of computation time, where *M* is the number of *enter* command sequences that cause leaks, in other words, there could be multiple leaking *x*s, so, *M* implies that there is at least one leak sequence exist. After the Grover iterations (calls to oracle *Leak_Detect*) were performed, one of the *M* sequences will be measured out with higher probability than the sequences that may or may not causing leak. The algorithm is a quadratic improvement over the *N*/*M* × O(*Leak_Detect*) calls performed by classical computer.

Since measuring from step (4) will render only one result, however, there could be cases that has no or multiple leak sequences exist, hence the result could be a random sequence i.e. mistakenly identified as a leaking sequence. To correct this inaccuracy, three methods can be applied:

a) A planned fake leak sequence x_f: *enter* (g_f, s_f, o_f) is assigned in between line 9 and 10 of *Leak_Detect* process in the *Safet_Test* algorithm such that the x_f will be detected *as* a leak command sequence in $\sqrt{N}$ time with high probability close to 100%, because it is the only leak sequence can be detected that makes *M* = 1. And if after several runs of the algorithm, the results repeatedly measured to be the same x_f, we can confidently determine that there are no other leak sequences besides the planned x_f.

b) To more precisely determine the number of leak command sequences, combine the Grover iteration *G* with **quantum counting** algorithm [14]. The method is to estimate the number of leak command sequences by quantum counting, which is an application of the phase estimation procedure to estimate the eigenvaluses $e^{i\theta}$ of Grover iteration *G*, which in turn enables determining an approximate number of leak command sequences *M*. The method allows us to decide whether a leak sequence even exists depending on the result number. The phase estimation circuit used for quantum counting is shown in Figure 6. The function of the circuit is to estimate θ to an accuracy approximate to 2^{-m} (note[2]).

[2] More accurate, *m* should be $m + \lceil \log(2+1/2\epsilon \rceil$ qubits .

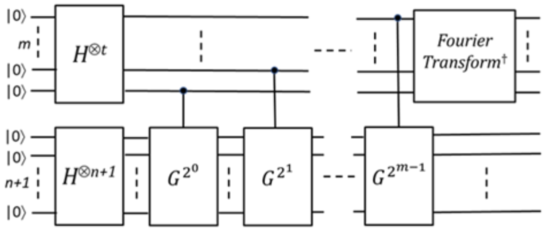


Fig. 6. Approximate quantum counting circuit for *G*

c) requires no additional process, but repeatedly running the algorithm enough times, then analyze the measured results. If there is no command sequence causing leak, any random sequence number will be measured with the same probabilities of all other sequence numbers. Such result indicates that there is no concentrate output of one particular leak sequence number meaning that the possibility of having a true leak sequence is low, however, this method is reliable only when the total number of actions, subjects and objects is large enough for the odd that getting a random result, which is true leak sequence is low. Table 3 compares the three methods.

Table 3. Comparison of testing methods for checking the existence of true leak sequences.

Checking methods	If true leak sequences exist	It no true leak sequence exist	Accuracy
(a) Plan fake leak sequence access x_f	Equal probabilities of getting true leak sequences x_{leak} and x_f	Fake leak sequence x_f has highest possibility being measured	Median
(b) Quantum counting by applying phase estimation	Number of solutions from phase estimation algorithm > 0	Number of solutions from phase estimation of algorithm ≈ 0	Hight
(c) No extra step required but run the algorithm enough times	High probability of a true leak x_{leak} sequence is measured	Equal probability for every sequence will be measured	Low (reliability increased by increasing the number of input (i.e. subjects and objects) sequences)

Table 3 shows that the more difficult in implementing the method (as ordered by methods *b*, *a*, and *c*) the more accurate result it will generate, unless depending on the number of possible sequences in method *c*, which if applied to a large number of total sequences (say no less than hundreds) then the accuracy might equal or better than method *a* and *b*, however, repeating the process of *c* method is not as efficient as the other methods. The detail algorithms and comparison of the three methods is interesting that worth to be discussed by their own topics, due to the limited space and to keep the discussion on focus, we only briefly introduce them in this paper.

5 Performance of safety check quantum algorithm

We can now summarize the performance improvement of safety test for a mono operational ABPS by comparing quantum to the classical algorithms. Assuming there is at least one leak sequence exist, by the quantum safety test algorithm, it will take $\sqrt{N} \times \mathrm{O}\ (Leak\ Detect)$ while classical safety test algorithm requires $N \times$ O (*Leak_Detect*) (for simplicity of demonstration, let's assume $M = 1$). The difference is in the order of $\sqrt{N}$ compared to $N = 2^{|A|\times|S|\times|O|}$, which is the 2's power of the number of actions |A| times the number of subjects |*S*| times the number of objects |*O*| managed by the ABPS system. For ICDSs or applications accessed by large number of subjects (users classified by attributes) to multiple number objects (devices classified by attributes), the quadratic difference is significant as shown in comparison listed in Table 4. Note that for the purpose of comparison, the O (*Leak_Detect*) is not counted, because both algorithms take the same polynomial time which does not affect the exponential difference.

Table 4. Computation time comparison of classical and quantum algorithms for ABPS safety test

Number of subjects times objects	Classical algorithm × O (*Leak_Detect*)	Quantum algorithm × O (*Leak_Detect*)
5	32	5.6568542494492 ≈ 6
10	1024	32
15	32768	181.0193359838 ≈ 181
20	1048576	1024
25	33554432	5792.61875148 ≈ 5793
30	1.073741824×10^{9}	32768
35	$3.4359738368\times10^{10}$	185363.8000474 ≈ 185364
40	$1.099511627776\times10^{12}$	1048576
45	$3.518437208883\times10^{13}$	5931641.601516 ≈ 5931642
50	$1.125899906843\times10^{15}$	33554432

The growth of computation time from 5 to 50 (number of subjects times objects) is about 3.5×10^{13} time for classical algorithm, and about 6×10^{6} time for quantum algorithm, obviously, the improvement of quadratic reduction by quantum algorithm allows the safety test to be reasonably performed.

An ICDS's device in general is accessed by only one public user class (subject with *public* attribute) plus one administrator (subject with *administrator* attribute) and limited actions available to manage the device, so, at minimum, two subjects can read and write (most common actions) to the object, thus, only require $2^{2\times2\times1}\times$ O (*Leak_Detect*) computation steps by classical algorithm for safety test. However, some ICDSs may have more than one device to be managed, so the access control policy is deployed from central service to individual device as described in Section III. In such cases, the ICDS's ABPS may apply one-size-fits-all access control policy to its devices, thus even with limited allowed actions, but has multiple number of administration subjects and device objects. Further even with a single device (object), it is not uncommon that an ICDS has more than tens even hundreds of subjects and objects. So the computation time for safety test is not practical by classical algorithm for these systems, but if instead use quantum algorithm, the difference is enormous even for small number of actions, subjects and objects as shown in Figure 7, the growth is measured in 1000 computation time per unit for up to 20 in comparison of classical and quantum algorithm. It shows that there is not much benefit using quantum algorithms if the number is less than 10, however, the difference is obvious when the number is greater. Note that the comparison is for detecting leak for only one action, if there are multiple actions involved, the computation time will increase even exponentially greater by |*A*|, which is the number of the actions under test as a factor of exponent.

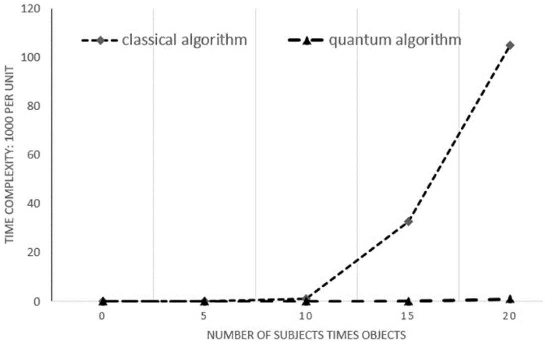


Fig. 7. Computation time comparison of quantum and classical algorithms for safety test

6 Conclusion

To determine the safety of a protection system is to find if there are privilege leaks from protected actions to unauthorized subjects of the system. HRU shows that for mono operational protection system, the computation time for the safety test is decidable, however take NP complete computation time, which is too expensive to perform for a system with large number of subjects and objects such as ICDS (e.g. IoT, RFID systems etc.) that applies attribute based access control (ABAC) model.

We demonstrate that an ABPS (protection system that applies ABAC model) such as ICDS can be simulated by an HRU access matrix and its matrix management functions. And adapted from Grover quantum search algorithm, we propose a quantum safety test algorithm, which determines the safety by returning a command sequence that will cause access leak for a mono operational ABPS. We conclude that if N equals to $2^{|A|\times|S|\times|O|}$ where |*A*| is the number of actions, |*S*| is the number of subjects, and |*O*| is the number of objects, and each of subject, object represent a set of attributes associate to them, the quantum algorithm for a mono operational ABPS requires computation steps $\sqrt{N}$ times the time required for classical leak detection process, compared to N times the time required for classical leak detection process, the quantum algorithm reduces the computation time quadratically. The saving is significant for ICDS or similar systems that its devices usually are accessed by large number of subjects with limited available actions. In addition to the quantum algorithm, three methods are explained to ensure that the test result are genuine instead of some random command sequence that does not but mistakenly rendered as a command sequence that causes access privilege leak.

References

1. Rouse, M.: Internet of Things (IOT), IoT Agenda, Tech Target, https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT (2019).
2. Voas, J., Kuhn, R., Laplante, P., Applebaum, S.: Internet of Things (IoT) Trust Concerns. NIST Cybersecurity White Paper (2018).
3. Siboni, S., Glezer, C., Shabtai, A., Elovici, Y.: A Weighted Risk Score Model for IoT Devices. In: SpaCCS 2019 International Workshops proceedings, pp 20-34. Springer, Atlanta. GA, USA (2019).
4. Hu. V. et. al: Guide to Attribute Based Access Control Definition and Considerations", National Institute Standards and Technology", NIST SP 800-162. (2014).
5. AXIOMATIC: Attribute Based Access Control (ABAC), https://www.axiomatics.com/attribute-based-access-control/.
6. Harrison, M. A., Ruzzo, W. L., Ullman, J. K.: Protection in Operating System, Communications of the ACM Magazine, Volume 19 Issue 8, pp 461-471 (1976).
7. Xu, Z., Li, X.: Secure Transfer Protocol Between App and Device of Internet of Things. In SpaCCS 2017 International Workshops, Proceedings, pp 25-34, Guangzhou, China (2017).
8. Skarmeta, A. F., Hern´andez-Ramos, J. L., Moreno, M. V.: A decentralized approach for Security and Privacy Challenges in Internet of Things, IEEE World Forum on Internet of Things, https://ieeexplore.ieee.org/abstract/document/6803122 (2014)
9. Dhillon, P., Singh, M.: Internet of Things Attacks and Countermeasure Access Control Techniques: A Review, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 7 pp. 1689-1698 © Research India Publications. http://www.ripublication.com (2019).
10. Mali, A., Darade, S.: Security and Privacy in Web-based Access Control in Internet of Things, Academia, https://www.academia.edu/28002646/Security_and_Privacy_in_Web-based_Access_Control_in_Internet_of_Things.
11. Maddison, J.: The Importance of Access Control for IoT Devices", SECURITYWEEK, https://www.securityweek.com/importance-access-control-iot-devices (2018).
12. Grover, L.: A fast quantum mechanical algorithm for database search. In Annual ACM Symposium on the Theory of Computation, page 212-219, ACM Press, New York (1996).
13. Grover, L. K.: Quantum mechanics helps in searching for a needle in a haystack, Phys. Rev. Lett, 79(2):325, 1997 arXive e-print quant-ph/9706033 (1997).
14. Nielsen, M., Chuang, I. L.: Quantum Computation and Quantum Information, Cambridge University Press (2000).

A Comprehensive Analysis on Multicast and Unicast Performance and Selection

Chen Shen[*][†], Chunmei Liu[‡], and Richard A. Rouil[‡]
[*]Associate, Wireless Networks Division, National Institute of Standards and Technology, USA
[†]Department of Physics, Georgetown University, USA
[‡]Wireless Networks Division, National Institute of Standards and Technology, USA
Email: [†]sc1951@georgetown.edu [‡]{chunmei.liu, richard.rouil}@nist.gov

***Abstract*—With the need to serve multiple users intended for the same content, especially in mission-critical applications, multicast has long been studied with evolving standards. Compared with its counterpart unicast, multicast has an apparent advantage of sending one copy instead of multiple copies. However, in Long Term Evolution (LTE) Multicast-Broadcast Single-Frequency Network (MBSFN), multicast does not support Multiple Input Multiple Output (MIMO) technology, which is one major technology that improves unicast performance significantly. Multicast also differs from unicast in other aspects that have major performance impacts, such as constructive signals and significant interference reduction, no retransmissions, less available subframes, extended cyclic prefix, and denser reference signals, to name a few. While almost all existing work focuses on a single factor and few addresses MIMO, in this paper we study multicast and unicast in detail with all these factors included, together with their integrated impact on performance. Profound analysis reveals that, contrary to what is commonly assumed in existing studies, multicast and unicast would not share the same modulation and coding scheme, but rather differ significantly in how efficiently they use resources. In addition, the balance among various factors mentioned above leads to a switch point where multicast or unicast outperforms, and the switch point changes upon system configurations and the performance metric of interest. Given that multicast configuration is semi-static in LTE, the results provide insightful guidelines in unicast or multicast deployment in serving user traffic. The work can also be easily extended to other Single-Frequency Network (SFN) based multicast technologies.**

***Index Terms*—Multicast, MIMO, throughput, file transfer time.**

I. Introduction

In wireless cellular networks, when there are multiple User Equipments (UEs) intending for the same content, the base station has two ways to deliver the content, unicast and multicast. In unicast, the base station sends multiple copies to the UEs, one copy per UE. Whereas in multicast, the base station sends one copy to all UEs of interest. Among multicast technologies and without loss of generality, in this paper we consider Long Term Evolution (LTE) Multicast Broadcast Single Frequency Network (MBSFN) due to that it is Single Frequency Network (SFN) based and has the potential to improve cell edge performance, which is critical to First Responders (FRs) to ensure coverage [1].

At first glance it appears that as long as there is more than one user, multicast would always outperform unicast in saving spectrum, since in case of N users, multiple copies in unicast would mean N times as much resource required as that in multicast. A closer look at this statement reveals that it is based on one assumption, that unicast and multicast use the resource as efficiently as their counterparts. This assumption is actually one implicit assumption made in most previous papers studying unicast and multicast, where it was typically assumed that one UE would share the same Modulation and Coding Scheme (MCS) for unicast and multicast [2]. However, this assumption does not hold in practice. On one hand, Multiple Input Multiple Output (MIMO) technologies are applied for LTE unicast, including both transmit diversity and spatial multiplexing, while for multicast, there is no transmit diversity and only one data stream is allowed. That is, MIMO technologies could allow unicast to use resources much more efficiently than multicast. On the other hand, multicast generates constructive signals and gains from interference reduction and diversity combining [3], which could favor multicast significantly in efficient resource utilization. The resulting resource efficiency for both multicast and unicast depends on multiple factors, such as UE distributions and channel conditions.

There are other factors that would lead to different performances, too [4]. One example is that in most cases multicast could use only six out of ten subframes (this constraint is relaxed conditionally to eight in later releases), while unicast could use all ten subframes. Other examples include extended Cyclic Prefix (CP) in multicast versus normal CP in unicast, and denser multicast reference signals (RSs) in resource grid. In addition, while unicast could use Hybrid Automatic Repeat Request (HARQ), there are no retransmissions in multicast, which means a lower target Block Error Rate (BLER) for reliability.

In this paper, we explore and compare unicast and multicast with all these factors embedded throughout the analysis. Since MIMO plays a significant role here, we start with antenna configurations in Signal to Interference plus Noise Ratio (SINR) and resource efficiency, the results of which distinguish unicast and multicast. We then consider the number of copies sent and proceed to throughput and file transfer time. We show that there exists a switch point in the number of users, where unicast or multicast outperforms. Also, the switch points differ upon different performance metrics selected and antenna configurations. The results provide guidelines in unicast or multicast deployment in serving user traffic, especially given that MBSFN configuration is semi-static as specified in 3GPP. To our best knowledge, this work is the first that considers these factors comprehensively. While the analysis and results are based on LTE MBSFN, they can be easily extended to other SFN based multicast technologies.

Shen, Chen; Liu, Chunmei; Rouil, Richard A. "A Comprehensive Analysis on Multicast and Unicast Performance and Selection." Presented at 2021 IEEE Global Communications Conference (GLOBECOM 2021), Madrid, ES. December 07, 2021 - December 11, 2021.

The following articles study various aspects of multicast technology in LTE. An overview of LTE Evolved Multimedia Broadcast Multicast Services (eMBMS) structure and mechanisms is given in [5]. The standards evolution of multicast broadcast technology in 3GPP is reviewed in [6]. In [7], the authors derive an MBSFN area formation algorithm with optimized overall throughput. In their modeling, the same per Resource Block (RB) throughput is applied for unicast and MBSFN, which is not practical in real cases due to MIMO in unicast and extended CP in MBSFN. We resolve this problem by including them in our modeling. With mobile edge computing and cache ability, the deployment of multicast can be even more flexible and efficient, which is modeled and examined in [8]. Our results could be extended with application of mobile edge computing facilities. Since MBSFN has an advantage in broadcasting but a disadvantage in MCS limitation from the worst SINR among UEs, its scheduling becomes an optimization problem on multicast grouping, which is discussed in [9]. Our analysis could be used to enhance the study by providing practical assumptions for modeling.

The rest of the paper is organized as follows. In Section II, we analyze the differences between unicast and multicast due to antenna configurations, in SINR and resource efficiency. In Section III we explore user performance and the resulting switch points in number of users. Lastly, we conclude our study in Section IV.

II. SINR, Rank, and Resource Efficiency

In the section we explore how efficient unicast and multicast utilize resources. We start with SINR then move to resource efficiency.

A. Network Design

The network studied has a typical hexagonal grid of 37 tri-sectored sites or 111 cells in total. Inter-site distance is 500 m. Base station transmission power used is 40 w, with transmitter (base station) and receiver (UE) heights of 32 m and 1.5 m, respectively. Band 14 with bandwidth of 10 MHz is applied. Consequently, each subframe has 50 RBs. Urban channel model defined by 3GPP [10] is used for path loss, and Claussen model for shadowing [11]. The small scale fading employed is the International Telecommunication Union (ITU) 'VehA' model with speed of 30 km/h. The center 7 sites, or 21 cells, define the area of interest where UEs are dropped, with other sites generating interference. For multicast, unless mentioned otherwise, the center 7 sites/21 cells form the MBSFN area, which is called 21-cell multicast in this paper.

TABLE I: Antenna Configurations

Number of Tx Antennas	Number of Rx Antennas	Unicast Transmission Mode
1	1	Single Input Single Output (SISO)
2	2	Open Loop Spatial Multiplexing (OLSM)
4	2	OLSM
4	4	OLSM
8	2	TM9
8	4	TM9
8	8	TM9

Seven antenna configurations are studied for both unicast and multicast, as listed in Table I. For unicast, transmission modes specified in 3GPP [12] are listed, which reflect MIMO technologies. For multicast, 3GPP [13] specifies no transmit diversity and single layer transmissions.

The link curves in [4], which consider the extended CP and denser RS pattern for multicast, are used for physical abstraction. For target BLERs, the typical value of 0.1 is selected for unicast, and 0.01 for multicast to compensate for no retransmissions.

B. Unicast

For LTE unicast, as discussed previously, spatial multiplexing is supported and data from multiple layers is jointly coded. For each transport block (TB) sent with MCS m and number of layers L, by using Mutual Information Effective SINR Mapping (MIESM) averaging [14] over all Resource Elements (REs) and layers, we can derive the Additive White Gaussian Noise (AWGN) equivalent post-equalization SINR as below:

$$\gamma(m, L) = f_m^{-1} \left[\frac{1}{L N_{\mathrm{RE}}} \sum_{l=1}^{L} \sum_{c=1}^{N_{\mathrm{RE}}} f_m(\gamma_l^c) \right], \tag{1}$$

where $f_m(\cdot)$ represents the Bit-Interleaved Coded Modulation (BICM) capacity with MCS m; N_{RE} is the total number of REs used; and γ_l^c is the post-equalization SINR over RE c for layer l.

Let (m_0, L_0) denote the optimal value of m and L where the UE maximum throughput is achieved. That is,

$$(m_0, L_0) = \arg\max{}_{(m,l)} throughput. \tag{2}$$

Then

$$\gamma_0 = \gamma(m_0, L_0) \tag{3}$$

is essentially the layer-level AWGN equivalent SINR that maps to the maximum throughput achievable by this TB.

For the seven antenna configurations under study, Figure 1 plots the cumulative density function (CDF) of the resulting γ_0. Note that the equivalent SINR is capped at around 25.76 dB. This is due to the limitation of the BICM mapping. The mapping is designed for SINR in range of [-20, 30] dB, and the highest equivalent SINR mapped back from BICM is 25.76 dB. When the input SINRs before mapping to BICM are above 30 dB, linear averaging is used instead, which we will see later in multicast equivalent SINR.

Figure 1 shows that the CDFs for cases 2x2 and 4x2 are similar which has around 5 dB gain over SISO. Additionally, the cases of 4x4 and 8x2 have comparable SINRs. In these cases a limited set of codebook-based precoders are used. We also observe significant gains by using eight Tx antenna over less Tx antennas. Furthermore, with the same Tx antennas, SINR advances with increasing number of Rx antennas. This is consistent with gains from multiple antennas.

For MIMO, in addition to SINR, another important factor to performance is number of layers, or rank, which maps to number of data streams. For SISO, the rank is always 1. To study rank for other antenna configurations, we drop UEs to

Shen, Chen; Liu, Chunmei; Rouil, Richard A. "A Comprehensive Analysis on Multicast and Unicast Performance and Selection." Presented at 2021 IEEE Global Communications Conference (GLOBECOM 2021), Madrid, ES. December 07, 2021 - December 11, 2021.

saturate the area of interest, and list in Table II the resulting percentage of UEs whose optimized ranks are above 1. It can be seen that although for configurations 2x2 and 4x2, the percentage is low, for other configurations (4x4, 8x2, 8x4, and 8x8), unicast does make good use of multiple data streams.

TABLE II: Percentage with Rank above 1

Antenna Config.	2x2	4x2	4x4	8x2	8x4	8x8
Rank >1	0.85 %	0.15 %	57.0 %	27.0 %	71.4 %	93.4 %

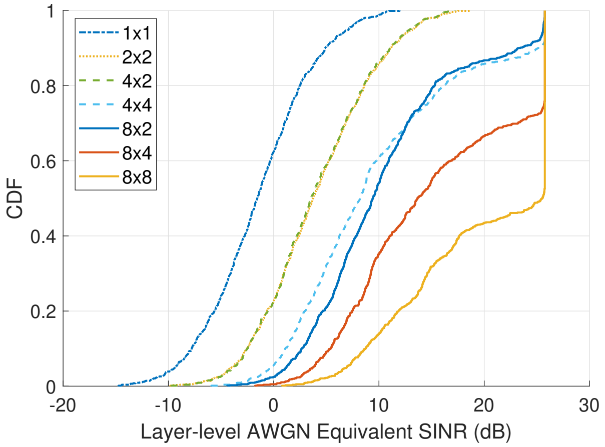


Fig. 1: Unicast Layer-level AWGN Equivalent SINR

SINR together with rank could reflect how efficient unicast uses resources. Instead of two metrics, a more straightforward way is to directly use one metric, resource efficiency, which is the sum of the number of bits each Orthogonal Frequency Division Multiple Access (OFDMA) symbol carries over all layers. Table 7.1.7.1-1 in 3GPP [12] lists the mapping from each MCS index to resource efficiency. Using this table, unicast resource efficiency can then be calculated from the Channel Quality Indicator (CQI) switching points in [4] and the optimal m_0 and L_0 from Eq. (2). For different antenna configurations, Figure 2 plots the resulting CDFs of the resource efficiency at the optimal point (m_0, L_0). As expected, the CDFs share similar trends as the SINR CDFs in Figure 1. In addition, 4x2 and 2x2 almost double the resource efficiency of SISO; and 8x8 almost double the ones of 8x2. This again indicates that unicast improves its performance significantly by taking advantage of MIMO technologies.

C. Multicast

Multicast MBSFN no longer employs transmit diversity and spatial multiplexing. However, multicast does have multiple cells transmitting constructive signals simultaneously to UEs, which results in significant interference reduction and diversity combining gain [3] [4]. Multicast also uses extended CP. By taking these into account, the post-equalization SINR for RE c can be modeled as:

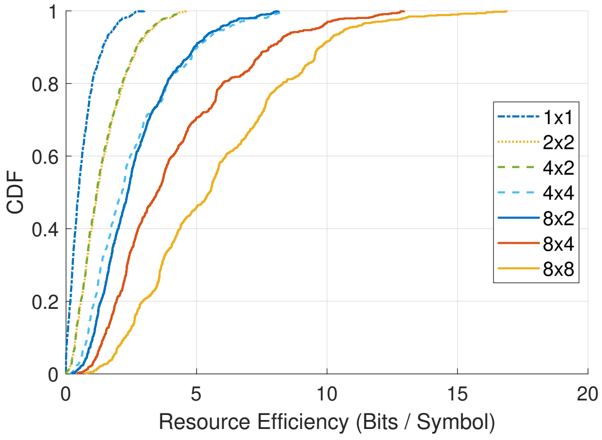


Fig. 2: Unicast Resource Efficiency

$$\gamma^c = \left(\sum_{i=1}^{N_{\mathrm{M}}} (1-\omega_i) P_i^c \| \boldsymbol{f}^c \boldsymbol{H}^{(i,c)} \mathbf{1}_{N_{\mathrm{Tx}}} \|^2 + \sum_{l=N_M+1}^{N} P_l^c \| \boldsymbol{f}^c \boldsymbol{H}^{(l,c)} \boldsymbol{W}^{(l,c)} \|^2 + \| \boldsymbol{f}^c \|^2 \sigma_N^2 \right)^{-1} \quad (4)$$

where N_M cells of total N cells are for MBSFN; ω_i represents the effective portion of signal from cell i within the extended CP; P is the signal power after path loss and shadowing but no small-scale fading; $\boldsymbol{H}$ stands for the frequency domain channel gains; $\boldsymbol{f}$ is for zero-forcing receiver; $\boldsymbol{W}$ represents the corresponding channel precoding matrix; and σ_N^2 serves as the thermal noise.

Similar to unicast, we apply MIESM averaging to obtain AWGN equivalent SINR:

$$\gamma(m) = f_m^{-1} \left[\frac{1}{N_{\mathrm{RE}}} \sum_{c=1}^{N_{\mathrm{RE}}} f_m(\gamma^c) \right]. \quad (5)$$

Different from unicast in Eq. (1), there is no longer averaging over number of layers. Also, there are values of γ^c higher than 30 dB before mapping to BICM. For these SINRs, as mentioned in Section II-B, linear averaging is applied instead of MIESM averaging.

Similar to unicast, let m_0 denote the MCS that achieves the maximum throughput. Then

$$\gamma_0 = \gamma(m_0) \quad (6)$$

is essentially the AWGN equivalent SINR that maps to the multicast achievable throughput. Note that by comparing Eq. (1) and (2) with Eq. (5) and (6), it is apparent that unicast and multicast differ in their optimal MCSs, and in number of data streams as well.

For the seven antenna configurations under study, Figure 3 plots the CDFs of the resulting γ_0. It shows that either increasing number of Tx antennas or number of Rx antennas or

Shen, Chen; Liu, Chunmei; Rouil, Richard A. "A Comprehensive Analysis on Multicast and Unicast Performance and Selection." Presented at 2021 IEEE Global Communications Conference (GLOBECOM 2021), Madrid, ES. December 07, 2021 - December 11, 2021.

both will result in higher SINR. Also because of the MBSFN gain from interference reduction and diversity combining [3], the SINR could achieve very high values, significantly above 20.5 dB required for the highest CQI [4].

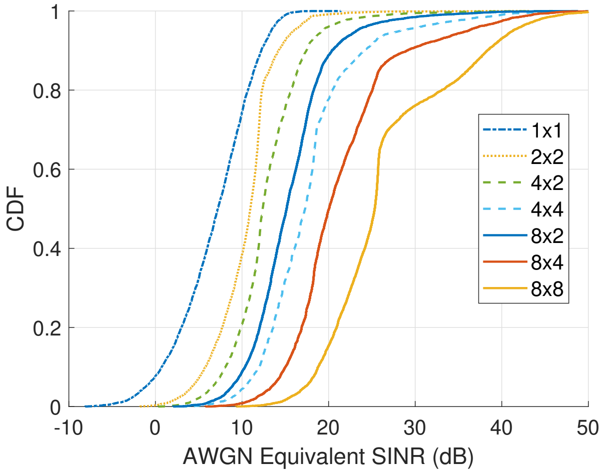


Fig. 3: Multicast AWGN Equivalent SINR

Unfortunately these high SINRs do not necessarily lead to throughput superiority because of the highest CQI cap [12] and lack of MIMO support. Figure 4 shows the resource efficiency, which demonstrates a more realistic achievable performance under various antenna configurations. Note that the extended CP, denser RSs and no retransmissions in multicast have been embedded into the calculation [4]. As expected, the relative positions of the CDFs follow the SINR CDFs in Figure 3. In contrast to unicast where very high resource efficiency can be achieved due to multiple layers (Figure 2), the multicast resource efficiencies are upper bounded by its single layer limitation, which is approximately 5.5 bits per symbol [12].

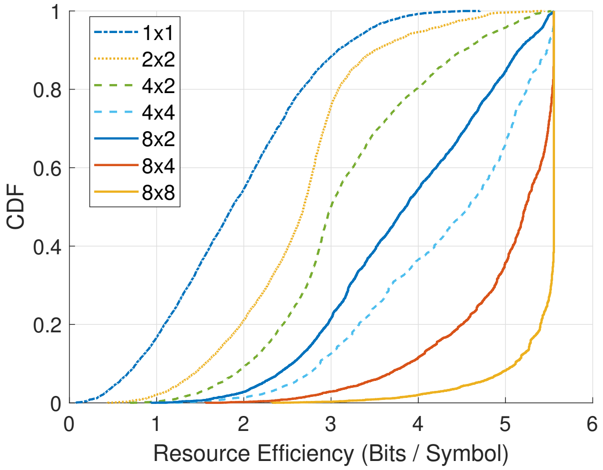


Fig. 4: Multicast Resource Efficiency

In LTE release 10, up to 8-layer downlink transmissions have been specified. Considering realistic device capabilities, the rest of the paper will concentrate on 8x4 antenna configuration with conclusive results for other configurations.

III. Switch Points

In this section we proceed to the impacts of the number of copies sent. Two performance metrics are selected, throughput and file transfer time. Consistent with 3GPP [15], multicast uses six subframes while unicast uses ten.

Due to its complexity and the amount of parameters involved, we use system level simulations for this analysis. The network considered is the same as Section II-A with 8x4 antenna configurations unless mentioned otherwise. For unicast, the typical proportional fairness scheduler is applied; whereas for multicast, all cells within the MBSFN area generate constructive signals by mapping all resources in a subframe to one TB and sending the TB to UEs simultaneously [15]. Additionally, to ensure that all UEs can correctly decode packets, multicast employs the lowest MCS among all UEs.

A. *Potential Throughput*

We start with potential throughput where there is one UE per cell, so that the impact from other UEs is removed. Consequently, both unicast and multicast send one copy. In addition, for unicast, this means that the entire resource, instead of a portion of it, is assigned to the UE of interest. For multicast, this means that the best MCS for this UE is applied, i.e., the restriction of the lowest MCS among all UEs is lifted. Essentially potential throughput is the highest throughput one UE can achieve at a particular location.

Figure 5 plots the potential throughput of one UE when it is located at each position within the entire center 7 sites, for both unicast and multicast. It shows that unicast potential throughput has a large spread, approximately (0 to 100) Mb/s. Contrarily, multicast potential throughput falls into a relatively small range (0 to 16) Mb/s. In other words, depending on its location, in unicast one user could experience excellent throughput in some areas, while suffer in some other areas; whereas in multicast the user experience is relatively consistent across the whole area, yet much lower than the high end of unicast throughput. The excellent unicast throughput is mainly due to unicast spatial multiplexing, which does not apply to multicast. Nevertheless, multicast improves throughput especially at cell edges, which comes from constructive signals and interference reduction [3]. This is consistent with our previous analysis in Section II on resource efficiency.

If we average potential throughput over the area, unicast gets around 30.5 Mb/s versus multicast 13.87 Mb/s. That is, on average unicast MIMO gains outweigh multicast gains. While unicast outperforms when there is a single UE, in the next subsection we investigate whether this holds with multiple UEs, where unicast sends multiple copies versus one copy in multicast.

B. *Actual Throughput and Switch points*

Potential throughput heatmaps provide a good view on the highest throughput one UE could achieve across the area. A

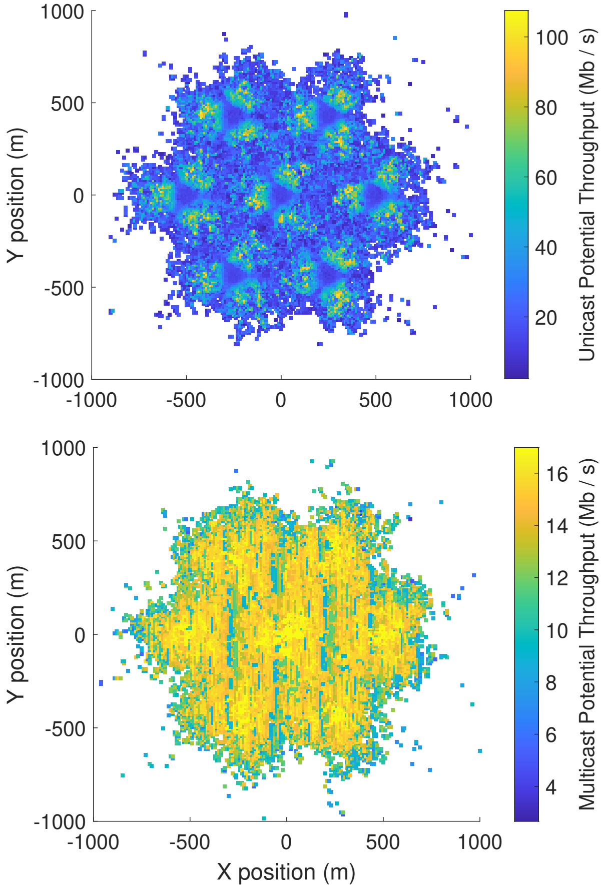


Fig. 5: Potential Throughput Heatmap

more realistic situation is when there are multiple UEs being served. We call throughput in this case actual throughput.

On one hand, in unicast, multiple UEs mean that more copies will be sent and the resource is no longer assigned to a single UE but rather split among multiple UEs. We hence expect that the more UEs being served, statistically the less resource each UE would get, and consequently the less actual unicast throughput. On the other hand, in multicast, it appears that multiple UEs do not affect throughput since one copy will be sent regardless of the number of UEs. However, recall that to ensure correct decoding, the lowest MCS among all UEs is used. Therefore, we also expect that the more UEs being served, statistically the lower MCS that would be used, and consequently the less actual multicast throughput.

As both actual throughputs decrease with increasing number of UEs, and unicast outperforms when there is a single UE as shown in potential throughput, it becomes interesting to see their relative performance with increasing number of UEs. For this purpose, we use Monte Carlo simulations with randomly dropped UEs in the area. The number of UEs per cell ranges from 1 to 10 with 100 simulation runs for each. The actual throughputs averaged over 100 runs are plotted in Figure 6, in blue and green. With increasing number of UEs, Figure 6 shows a sharp drop in unicast actual throughput, much faster than multicast. Plus the higher initial unicast actual throughput with one UE, together they lead to a switch point. Before the switch point, unicast outperforms, while beyond the switch point, multicast outperforms. In this particular setting, the switch point is 4 UEs per cell.

Moreover, in unicast the resource each UE gets is almost inverse to the number of UEs, statistically. It is thus expected that the unicast actual throughput is close to a reciprocal function, and that the throughput would converge to zero. On the contrary, for multicast, the limitation comes from the lowest MCS. It is thus expected that multicast actual throughput would eventually converge to the throughput that maps to the lowest MCS, excluding those UEs falling out of coverage.

Since the MBSFN area size affects performance [16], actual throughput with MBSFN area size of one cell is also plotted in Figure 6 in yellow. Similar to the 21-cell multicast and as expected, the one-cell multicast actual throughput also drops, and there also exists a switch point. Interestingly, although one-cell multicast gets less multicast gains from multiple cell transmissions, it has slightly higher actual throughput than 21-cell multicast, and this higher throughput also leads to a leftward shift of the switch point. This is because the lowest MCS employed is among all UEs within the entire MBSFN area. With the same number of UEs per cell, the more cells in the MBSFN area, the higher probability of a lower lowest MCS, hence a lower actual throughput in 21-cell multicast.

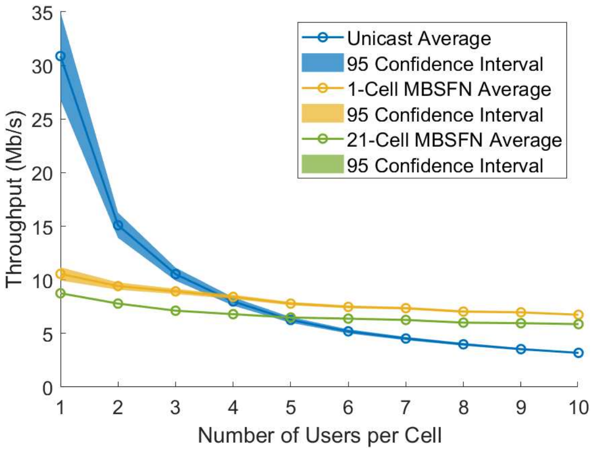


Fig. 6: Throughput Switch Point

Note that the switch point discussed above is based on average actual throughput. While it holds statistically, there exist individual cases that do not follow the switch point. We call these cases irregular cases. For example, when the number of UEs per cell is 2, which is before the switch point, unicast outperforms statistically. However, out of our simulation runs, around 18 % of the runs have multicast outperforming. These cases count towards the irregular cases. Table III lists the

Shen, Chen; Liu, Chunmei; Rouil, Richard A. "A Comprehensive Analysis on Multicast and Unicast Performance and Selection." Presented at
2021 IEEE Global Communications Conference (GLOBECOM 2021), Madrid, ES. December 07, 2021 - December 11, 2021.

percentage of irregular cases for different numbers of UEs per cell. As expected, the percentage rises as the number of UEs per cell gets closer to the switch point, and it can be as high as almost 50 %. Further digging into the data shows that its penalty could be as high as about 2 Mb/s in throughput or about 30 % in throughput percentage. This means that while we could use switch points as a guideline in selecting unicast or multicast to serve traffic, further investigation into irregular cases could help ensure the performance of individual cases.

TABLE III: Percentage of Irregular Cases

UE per Cell	1	2	3	4	5	6	7	8	9	10
%	7.3	18.2	31.1	49.1	36.6	26.1	19.2	14.9	10.6	7.9

The above analysis focuses on 8x4 antenna configuration. As discussed in Section II, different antenna configurations lead to different resource efficiencies for both unicast and multicast, hence different relative actual throughput and switch points. Table IV lists the switch points for the seven antenna configurations studied. Interestingly, on one hand, in unicast higher number of antennas leads to higher MIMO gains, and hence higher unicast throughput and potentially larger switch point. On the other hand, in multicast higher number of antennas also leads to higher resource efficiency as in Figure 4, and hence higher multicast throughput and potentially lower switch point. The numbers in Table IV show that these two effects balance differently under different antenna configurations. Consequently, switch points differ under different antenna configurations, but there is no obvious pattern in switch points versus antenna configurations.

C. File Transfer Time and Switch Points

In addition to throughput, another typical performance metric is file transfer time used for small file transfer. It is the duration from the start of the file transfer to the time the last UE receives the file. We simulate three file sizes with 200 repetitions for each. The resulting average file transfer time is shown in Figure 7.

It can be noted in Figure 7 that unicast file transfer time increases almost linearly with the number of users per cell, whereas multicast file transfer time increases but at a much slower rate. The underlying reasons are the same as those in previous throughput analysis, that more UEs in unicast means less resource for each UE and in multicast means lower MCS employed. Also, with small UE numbers, unicast has shorter file transfer time. Together with the faster increasing rate of unicast, a switch point is formed. In cases of the three file sizes simulated, the switch points are all around three UEs per cell. Recall that the throughput switch point is between 4 and 5. This difference in switch points is due to different amount of data transferred. In case of actual throughput, different UEs will have different amount of data transferred due to the proportional fairness scheduler; while in case of file transfer time, all UEs will have the same amount of data transferred. That is, with the same amount of data transferred, multicast has larger relative gains.

The switch points for all seven antenna configurations are listed in Table IV. Similar to the throughput case, switch points differ under different antenna configurations, but there is no obvious pattern. Note that compared with throughput, all switch points shift lower, which is consistent with the 8x4 configuration discussed above.

TABLE IV: Switch Point

Antenna Config.	1x1	2x2	4x2	4x4	8x2	8x4	8x8
Average Actual Throughput	10	10	4	3	5	4	5
File Transfer Time (0.05 Mb)	1	3	1	2	3	3	3

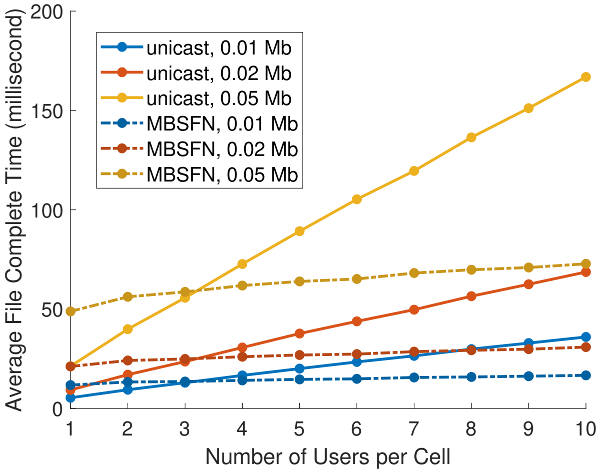


Fig. 7: File Transfer Time Switch Point

IV. Conclusion

In this paper we noted that MIMO technologies provide significant gains for unicast transmissions, whereas multicast gains from sending one copy to multiple users instead of multiple copies. We also noted other major factors on performance such as multicast constructive signals and less available subframes. With all the factors included in analysis, we first studied resource efficiency and showed that higher number of antennas improves resource efficiency not only for unicast, but also for multicast. We then explored user experience in terms of throughput and file transfer time. Detailed analysis revealed that there exists a switch point in number of users, where unicast or multicast outperforms.

Although the analysis and results are based on LTE MBSFN, they can be easily extended to other SFN based multicast technologies. Additionally, while the switch points can provide guidelines in selecting unicast or multicast in serving traffic, there exist irregular cases that do not follow switch points. Our next step is to extend the work to other multicast technologies, and to investigate irregular cases to ensure performance of individual cases.

References

[1] 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*, 2008, pp. 1–5.

[2] A. Alexiou, C. Bouras, V. Kokkinos, A. Papazois, and G. Tsichritzis, “Efficient mcs selection for mbsfn transmissions over lte networks,” in *2010 IFIP Wireless Days*, 2010, pp. 1–5.

[3] C. Liu, C. Shen, J. Chuang, R. A. Rouil, and H. Choi, “Evaluating Unicast and MBSFN in Public Safety Networks,” in *IEEE PIMRC 2020 - 2020 IEEE International Conference on Communications*, Aug. 2020.

[4] 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.

[5] A. Urie, A. N. Rudrapatna, C. Raman, and J. Hanriot, “Evolved multimedia broadcast multicast service in LTE: An assessment of system performance under realistic radio network engineering conditions,” *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 57–76, 2013.

[6] A. Sengupta, A. Rico Alvarino, A. Catovic, and L. Casaccia, “Cellular Terrestrial Broadcast—Physical Layer Evolution From 3GPP Release 9 to Release 16,” *IEEE Transactions on Broadcasting*, vol. 66, 2020.

[7] C. Borgiattino, C. Casetti, C. . Chiasserini, and F. Malandrino, “Efficient area formation for LTE broadcasting,” in *2015 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2015, pp. 202–210.

[8] R. Hwang, C. Wang, J. N. Hwang, Y. Lin, and W. Chen, “Optimizing Live Layered Video Multicasting over LTE with Mobile Edge Computing,” *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2020.

[9] S. Feng, C. Liu, C. Shen, H. Choi, and R. A. Rouil, “An Effective and Efficient Dynamic eMBMS Multicast Grouping Scheduling Algorithm in MBSFNs for Public Safety Scenarios,” *IEEE Access*, vol. 8, 2020.

[10] 3GPP TS36.942, “Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios,” 3GPP, , Jul. 2018.

[11] 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.

[12] 3GPP TS36.213, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures,” 3GPP, Standard, Jan. 2019.

[13] 3GPP TS23.246, “Multimedia Broadcast/Multicast Service (MBMS) Architecture and Functional Description,” 3GPP, Standard, 9 2019.

[14] Z. Hanzaz and H. D. Schotten, “Analysis of effective SINR mapping models for MIMO OFDM in LTE system,” in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2013, pp. 1509–1515.

[15] 3GPP TS36.101, “Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception,” 3GPP, Standard, 6 2020.

[16] C. Shen, C. Liu, R. A. Rouil, and H. Choi, “Study of Multicast Broadcast Single Frequency Network Area in Multicast Communications,” in *IEEE ICSPCS 2020 - 2020 IEEE International Conference on Signal Processing and Communication Systems*, Dec. 2020.

Advanced Sensing Development to Support Robot Accuracy Assessment and Improvement

Guixiu Qiao*, Senior Member, IEEE*

Abstract— **Robots can perform various types of automated movements in the workspace. In recent years, robot applications have been expanded to a much wider scope, including robot machining, robot assembly, robot 3D printing, robot inspection, etc. Many of these applications require robots to have higher absolute accuracy compared with conventional robot part handling and welding. The capability to assess a robot's accuracy, and further, improve accuracy becomes important. In this paper, an advanced sensor and an accompanying methodology are developed that enable manufacturers to perform accuracy assessment to improve their robot systems. A smart target (patent pending) is developed at the National Institute of Standards and Technology (NIST). This sensor is integrated with a vision-based measurement instrument to perform high accuracy measurements of six-dimensional (6-D) information (x, y, and z position, roll, pitch, and yaw orientation) for a moving robot arm. The smart target is motorized. It can constantly rotate toward the vision-based measurement instrument to maximize its line-of-sight. A use case is presented to demonstrate the accuracy degradation assessment by using the smart target on a Universal Robot.**

I. Introduction

Using standard six-degree of freedom (DOF) industrial robots for precision applications represents a significant market potential [1-3]. The main drawback of industrial robots is that the accuracy may not satisfy the requirements of precision applications. For example, in the automobile and aerospace industries, scanners are widely used in reverse engineering or part inspection for quality control. Scanners are mounted on the robot arm and move around with the arm to measure a large part or panel. The robot's rotation and position information are used for scanner data registration. Generally, the registration accuracy needed for scanner data is about 0.05-0.1 mm in position and 0.1- 0.5 degrees in orientation [1]. The most exacting commercial robot cannot reach the accuracy within the range. There need to be solutions to assess robot accuracy, perform improvements by calibration, or add an external tracking system that provides accurate position and orientation feedback.

Calibration can improve a robot's accuracy if errors are repeatable. There are two types of errors: environment-dependent errors and robot-dependent errors [1]. Environment-dependent errors come from robot setup or the changes in the environment, for example, the stability of the floor where the robot is mounted, or the temperature changes that influence the accuracy. Robot-dependent errors can be divided into geometrical errors, non-geometrical errors, and system errors [4]. Geometrical errors deal with the imperfection of linkage parameters. Denavit-Hartenberg (DH) parameters are the most popular correction factors for geometrical error calibration [2]. Gear backlash error is usually classified as geometrical errors. Non-geometrical errors are related to gravity deformation, joint compliance, and hysteresis in gear transformation. System errors result from improper tool calibration or faulty sensor measurements [1]. Robot calibration primarily deals with robot-dependent errors. Before calibration, the robot is fully warmed up and the setup is carefully checked to eliminate the environment-dependent errors. When no force is applied, a robot can be calibrated with a complex model that contains both geometric and non-geometric errors to achieve the improvement in static positioning [5]. But when dynamics is involved or force is applied, for example, in robot machining operations, errors become non-repeatable and process-related. A real-time feedback approach is needed to compensate for the robot trajectory based on some sensor information.

Sensors used for static calibration or real-time feedback can be divided into absolute measurement and relative measurement. Absolute measurement captures the actual positions of the robot tool center position (TCP). The measurement is based on a world coordinate system, for example, theodolite measurement devices [6], laser trackers [4], and probing coordinate measuring machines (CMM) [5]. The measuring equipment can provide high-accuracy measurements. However, they are expensive and require skilled personnel. Relative measurement captures the relative position or orientation of the TCP. These measurements have constraints with single-point, plane, sphere, or distance to derive some of the kinematic parameters [7]. Relative measurements have limitations that work only for certain local areas. Because the measurement is constrained, for example, in a plane, the corrections are usually designed to correct a robot's operation in a similar constrained operation. In this paper, we focus primarily on the absolute measurement sensors.

The absolute measurement sensors output 6-D measurements under the equipment's world coordinate system. The measured position and orientation information can be used for robot accuracy assessment, or calibration, or real-time feedback. Besides being expensive, existing absolute measurement sensors have a challenge in dynamic measurement. For example, a laser tracker system usually mounts three reflective targets on a robot's TCP. The laser tracker will measure each target in sequence and output the center of the three targets as (x, y, z) positions in the tracker's

Guixiu Qiao is with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (phone: 301-975-2865; fax: 301-990-9688 e-mail: guixiu.qiao@nist.gov).

Qiao, Guixiu. "Advanced Sensing Development to Support Robot Accuracy Assessment and Improvement." Presented at 2021 IEEE International Conference on Robotics and Automation (ICRA 2021), Xian, CN. May 30, 2021 - June 05, 2021.

coordinate system [8, 9]. The three points can create a local TCP coordinate that represents the position and orientation of the TCP. When the robot arm moves, by remeasuring the three points, the translation and rotation of the TCP are captured with high accuracy. However, the robot needs to be stationary when the tracker is taking measurements of the three targets. Similarly, Lightcap [10] presented a method of using CMM to improve robot position accuracy. An apparatus with three tooling balls is mounted on a robot's TCP to be measured by a CMM. The CMM measurement process is slow, not suitable to capture dynamic robot movements.

Compared with these high-precision systems, optical tracking systems can provide 6-D measurements as well, but with lower accuracy [11]. Optical trackers use infrared (IR) cameras and infrared flash to highlight the reflective spherical targets and measure the sphere centers. The advantage of this type of system is that the vision-based system can capture all sphere centers simultaneously in one snapshot. Combined with a high-speed camera, the optical tracking system can measure at a high frame rate for dynamic movements. The challenge lies in measurement accuracy. The reflective target is captured as a dot in the IR camera. Different distances, ambient light, and target array orientations can influence the measurement uncertainty, particularly the orientation accuracy.

In order to perform high accuracy dynamic measurements of the robot TCP, a new sensor and a methodology were developed at the National Institute of Standards and Technology (NIST) to support robot accuracy assessment and improvement. This work is part of NIST's Prognostics and Health Management (PHM) for Reliable Operations in Smart Manufacturing project. The goal of this project is to develop and deploy 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 [12].

This paper is organized as follows: Section II describes the smart target sensor and methodology developed at NIST used in robot accuracy assessment and improvement. In Section III, the hardware and software of the smart target sensor are presented. A use case is presented in Section IV to show the accuracy assessment for a Universal Robot.

II. Robot Accuracy Assessment Methodology

A robot can approach a point in space from different directions using different poses. Accordingly, the accuracy varies with different approaches. It is impractical to take unlimited measurements to assess a robot's accuracy or perform accuracy improvement (for example, calibration). A methodology is needed to provide an efficient solution.

A methodology developed at NIST is shown in Fig. 1. It consists of multiple modules, including advanced sensing development, measurement plan and modeling method, taking measurements, data analysis, PHM remedy development, and control system feedback. First, an advanced sensor is developed (NIST patent pending) to measure the x, y, z, roll, pitch, and yaw of a moving robot's TCP. This work is detailed in Section III. In the next module in Fig. 1, a measurement plan is created. The pose of the robot arm cannot be picked arbitrarily. A measurement plan needs to be suitable for parameter identification of the robot error model. To satisfy this purpose, a fixed-loop motion is designed [13]. As shown in the left picture of Fig. 2, a set of poses is created within the robot workspace. These poses are distributed evenly in the robot Cartesian space. The even distribution in the entire workspace is to guarantee all the rigidity conditions - far, near, high, and low are considered. Additionally, the set of robot motions needs to be distributed evenly in the robot's joint space as well. The robot modeling method needs to identify parameters that can characterize the robot errors. When data is collected that is evenly distributed along the joint angles, the identification of the error model will have less chance to overweight or underweight some areas. These fixed-loop motions need to pass validation and collision checks [13]. The right picture in Fig. 2 shows the fixed-loop motion generated for a Universal Robot (UR5) after passing the checks. Because the robot is mounted on a table, only poses above the table are kept.

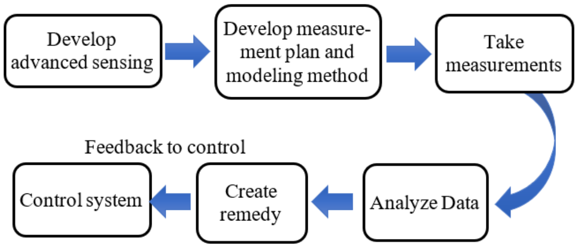


Figure 1. Robot accuracy assessment methodology

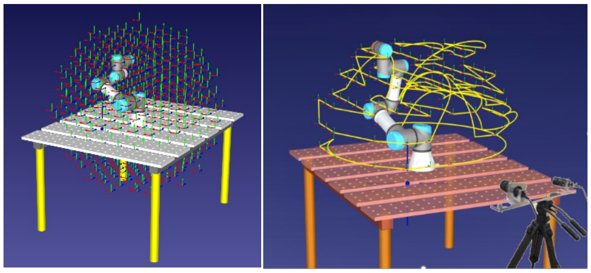


Figure 2. Measurement plan development

After the fixed-loop motions are generated, the program is loaded to a robot. The smart target is mounted on the robot TCP. The robot arm moves based on the designed fixed-loop. A vision-based measurement instrument is placed on the floor opposite to the robot arm. The vision-based measurement system measures the smart target on the TCP, outputting the x, y, z, roll, pitch, and yaw of the robot TCP movement. The Fig. 2 right picture shows a vision-based system capturing a UR5's fixed-loop motion.

Measurement data is sent to the data analysis module to identify the parameters of the error model and output the result of the accuracy assessment. The measured data can also be used for robot calibration to improve robot accuracy. The detailed modeling and parameter identification algorithms are documented in [14]. Then remedy suggestions are given for PHM purposes, for example, some minor adjustments are needed, or calibration is preferred. Feedback is sent to the control system for changes. This paper focuses on advanced sensing development, including design theory, hardware, and software to support the 6-D information capture for robot accuracy assessment and improvement.

III. Advanced Sensing Development

The goal of the advanced sensing development is to develop a non-contact 6-D measurement system, with high speed (at a minimum of 30 Hz) and high accuracy (within 0.1 mm). High speed enables the capture of the dynamic movement of the robot arm. High accuracy satisfies the requirement for robot calibration and accuracy assessment. Additionally, being cost-effective is another consideration for the development of advanced sensing in industrial applications.

The advanced sensing system developed at NIST consists of a smart target and a vision-based instrument, as shown in Fig. 3. The smart target is mounted on the robot TCP. The vision-based instrument performs non-contact measurements on the smart target. The smart target uses features to represent the 6-D information. An efficient way to represent 6-D information is to construct a coordinate frame. A coordinate frame carries the position and orientation information and can efficiently calculate transformations using a matrix. Therefore, the design theory of the smart target is to provide measurable features for the vision-based instrument. The vision-based instrument can capture features to construct a coordinate frame in a snapshot.

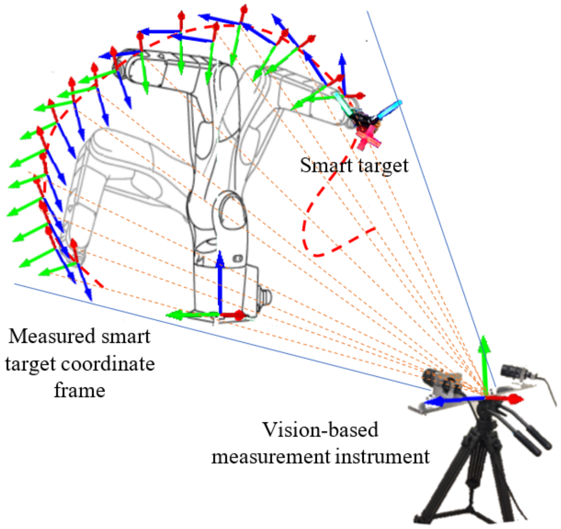

Smart target
Measured smart target coordinate frame
Vision-based measurement instrument

Figure 3. Smart target measurement system

There are many ways to build a coordinate frame. One simple way is to define an origin and two axes directions. The third axis is naturally perpendicular to the other two. For optical tracking systems mentioned in Section I, three spheres are used as the target to create a coordinate frame. One of the sphere centers is used as the origin. The other two are used to define axes. Unfortunately, the uncertainty of the axis direction is high because it is defined by only two points. Moreover, the distance between the two points is short with the limitation of target size, thus, enlarging the error of line-direction calculation. Additionally, there are line-of-sight problems. When the targets rotate with the robot arm, one sphere may block the view of others. For optical tracking systems, users cannot guarantee a target will always be facing the measurement instrument. Some angles may not be sensitive or measurable.

The advanced design of the smart target uses line features. The design enables a more accurate feature measurement than existing targets in the market, particularly in orientation accuracy. It has a mechanism to always rotate toward the measurement instrument. The smart target is robust to ambient light influences, which is a big hurdle in many manufacturing environments. The system will also be more cost-effective compared with a laser-based system that has a line-of-sight problem.

A. Hardware development for the smart target

The smart target is developed using line features. The line features are created by fixed-wavelength light pipes. The smart target also has two high-precision rotary gimbals, as shown in Fig. 4. Instead of using point features, the smart target uses line features to construct the coordinate frame. Light pipes are machined precisely in a cylindrical shape and the surface is specially treated to enable the evenness of lighting. Line features are extracted by measuring the light pipes with the vision-based instrument. An origin is created by intersecting the two feature lines capturing from the cross light pipe. The other two light pipes with different colors provide the other two axes directions. By using a cylindrical shape, line features are constructed using hundreds of points. The accuracy is greatly improved by the best-fit of hundreds of points instead of using two spherical center points. The origin is defined by the intersection of two lines, which is more accurate than using a single spherical center. Thus, the smart target can achieve higher accuracy than traditional targets.

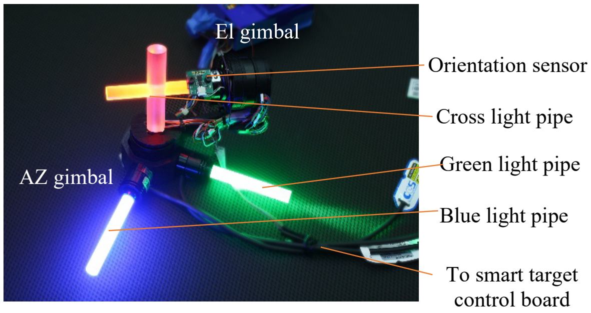


Figure 4. Smart target hardware design

There are two high-precision rotary gimbals in azimuth (AZ) and elevation (EL) directions, driven by two motors. An orientation sensor is mounted on the EL gimbal. The initial orientation of the smart target is set up to face to the measurement instrument. When the robot arm rotates, the orientation sensor on the smart target will transfer the detected rotation angles to the smart target control board. Then the AZ and EL motors start to rotate back to maintain the facing pose of the smart target toward the measurement instrument. Three fixed-wavelength light powers (red, green, and blue) are used to light up the light pipes. The different colors benefit the algorithms for feature identification. The smart target is scalable to bigger-sized industrial robots, or smaller-sized medical robots.

B. *Software development for feature extraction and 6-D information construction*

The dynamic measurements of a robot's motion require high speed and high accuracy. An optimized algorithm for GPU (Graphics Processing Unit) calculation is developed to satisfy the requirements. Computations for 6-D information construction include image un-distortion, feature extraction, 2D to 3D construction, and 6-D information output. A software library (software development kit, SDK) is developed for obtaining 6-D information from image pairs obtained from the vision-based measurement instrument. The SDK requires sub-pixel level accuracy (0.1 pixels) of feature extraction, critical for high precision measurement.

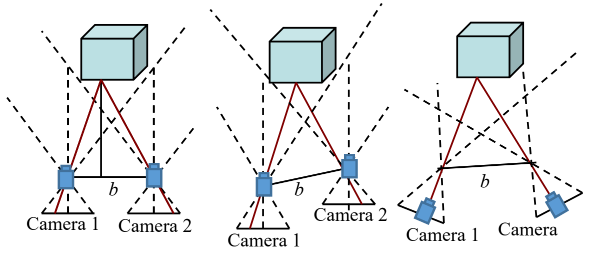
Many factors may create errors when matching two corresponding points in a pair of images, for example, when the images are blurry, or using local instead of global correspondence maximum, etc. The 3D measurement accuracy in the camera plane (also called the XY plane) is shown in Eq. (1). The accuracy aligned with the viewing direction (axis Z) is shown in Eq. (2).

$$\sigma_{XY} = \frac{q_D}{\sqrt{k}} m\delta_{x'}, \qquad (1)$$

$$\sigma_z = \frac{h}{b}\frac{q_D}{\sqrt{k}} m\delta_{x'}, \qquad (2)$$

$\delta_{x'}$ is the uncertainty of an image measurement. It is can be demonstrated by the subpixel accuracy achieved during calibration. Our design can achieve less than 0.1 - 0.2 pixels residual error of $\delta_{x'}$ from calibration results (calibration is not detailed in this paper). The parameter m is the image scale number. A camera system's m is the ratio of object distance h to the principal distance *c.* The principle *c* is lens focal length with the addition of an extension to achieve sharp focus. The larger h and smaller c created less accuracy. k is the mean number of target images per photo. The general value for k is usually set one. When more camera stations are added or using multiple exposures, the k value can be increased. The parameter b is the base distance of the two stereo cameras. The larger the base distance, the smaller σ_z error can be achieved. The parameter q_D is a design factor. It is related to the design of stereo camera configurations [15].

Fig. 5 shows three typical stereo camera configurations. The configuration in Fig. 5a) has two parallel cameras. It is closer to human eyes with nearly parallel viewing directions. The advantage of this configuration is that the dual cameras have the same scale factor. The disadvantage is the design factor q_D is the worst. It will enlarge σ_{XY} *and* σ_z error. The Shifted camera configuration in Fig 5b) also has parallel viewing directions but with a shift in depth. This configuration affects the width of the image frame. It also requires different scales and post-processing to shift the images. In the convergent configuration in Fig. 5c), the cameras are toed-in. This configuration has more overlaps for the two-camera view to provide a larger field of view for the measurement system. Also, the design factor is the best among the three configurations. However, because two cameras have more diversity in view, finding common points between the two images is more challenging. The difficulties are on the matching algorithms. Our design uses the convergent configuration to achieve a larger field of view and better accuracy.



a) Normal configuration b) Shifted configuration c) Convergent configuration

Figure 5. Stereo camera configurations

Fig. 6 shows one pair of smart target measurements from the stereo cameras.

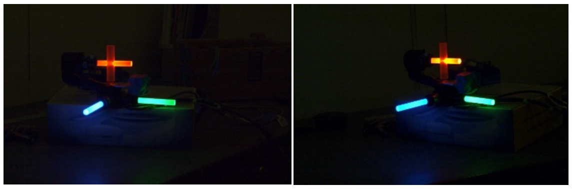


Figure 6. Smart targets images from left and right cameras

The feature detection algorithm is written in C++. The preliminary detection procedure is as follows.

Step 1. Image un-distortion

For smart target 6-D vector detection, the first step is to load the calibration results and use the parameters to undistort the images. A vision-based measurement instrument needs two calibrations – distortion calibration and stereo calibration. Distortion calibration corrects the lens distortions for individual cameras. Stereo calibration finds the relative physical positions of the two cameras. Generally, these two calibrations are mixed. Performing the full camera calibration takes time. After a measurement instrument is shipped or used for a while, the physical positions of the two cameras are feasible to change compared with camera lens distortion. In our research, the stereo calibration is separated from the distortion calibration. The stereo calibration is a much simpler process compared with the full calibration process with distortion calibration. When a test is performed on-site, the separated calibrations save time in system preparation. The calibration algorithms and procedures are detailed in the paper [14].

Step 2. Canny edge detection

In our preliminary approach, the color images are converted to grayscale images before edge detection. Later segmentation of color is developed for faster image processing. The RGB (red, green, and blue) colored images are analyzed using color thresholding to separate for the detection of red, blue, and green light pipes. Multiple edge detection algorithms were tested. Finally, the Canny edge detector is used to detect edges on all three color channels [15].

Step 3. Hugh line detection and clean up

The Hugh line detection algorithm is implemented to detect lines [16]. However, redundant or even false detections may occur. A cleanup algorithm is developed to address this problem. Fig. 7 shows the detected line after cleaning up.

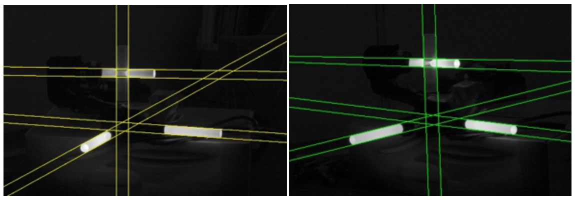


Figure 7. Detected lines from left and right cameras after cleanup

The final four vectors are shown in Fig. 8.

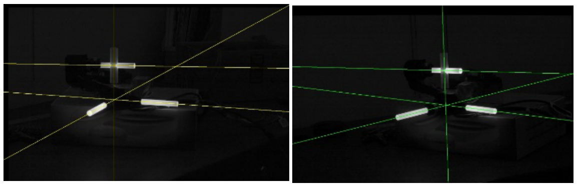


Figure 8. The final 4 vectors cameras after cleanup

Step 4. Epipolar geometry estimation

The construction of the 3D vectors from the 2D on stereo image pairs is based upon epipolar geometry [16]. Given the essential matrix and fundamental matrix from stereo calibration results, a vector v and its corresponding vector v' on the left and right camera, and a point x from the left camera on one of the vector, we can calculate the epiline on the right camera, which is shown in Fig. 9. The corresponding point of x on the right camera is the intersection between the epiline and the vector v'. Given the corresponding points x and x', we can reconstruct a 3D point X.

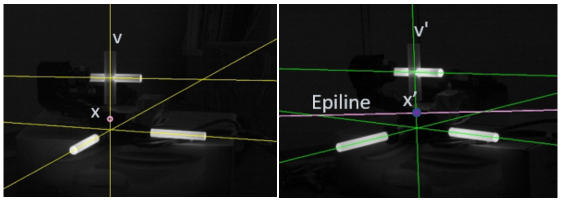


Figure 9. using epipolar geometry to calculate correspondence from two cameras

To improve the speed and robustness of detection, the stereo rectification of the pairs of the camera images is introduced. Rectification consists of aligning the image points in both the left and right images to a common global plane. In stereo rectification, the images are transformed so that epipolar lines are merged along with horizontal scan lines of the image. Given the rectified images, the 3D reconstruction is directly correlated to the disparity of the image pairs.

Step 5. Smart target 3D coordinate construction

We are looking for the following measurements: the intersection of the red cross, and the intersection of the blue and green bars, as shown in the white cross markers in Fig. 10. and the 3D vector of the green bar and blue bar. The final smart target coordinate is constructed using the cross center of the red cross light pipe as the origin. Two axis-directions come from vectors of the green and blue light pipe.

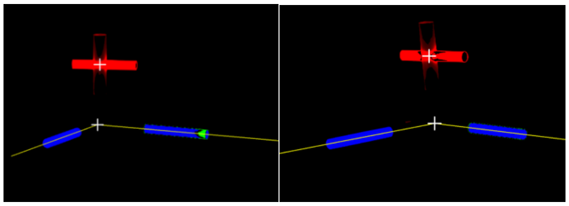


Figure 10. Final detected vectors (yellow lines) and coordinate centers (white cross)

C. Speed analysis

Using a Nvidia Quadro RTX 5000 GPU and 2.4GHZ INTEL Xeon W-10885M CPU (Central Processing Unit), we have the preliminary estimation of the calculation speed for each step as shown in Table 1.

Table 1. Speed analysis (in millisecond (ms))

Upload to GPU time	1.5 ms
Split channel and thresholding	5 ms
Un-distortion/remap	5 ms
Edge detection	7 ms
Line detection	3 ms
Line cleanup	0.05 ms
Download from GPU time	3.5 ms
Total Time	25 ms

Currently, the camera can run at 125HZ. The image processing takes 25 ms to finish processing one pair of images. Efforts continue in algorithm improvement and CUDA (Compute Unified Device Architecture) optimization. And the time consumption is based upon the current CPU/GPU combination.

IV. Use Case Analysis and Conclusion

A use case was developed using a UR5 robot to assess the robot accuracy degradation when payload, speed, and temperature changes are considered. As shown in Fig. 11, a smart target was mounted on the last link of the UR5 robot. The vision-based measurement instrument was placed on the floor. The smart target was initialed to face to the measurement instrument. When the robot moved, the smart target maintained the same facing direction. For the test, the same program was repeated to drive the robot’s movement at different conditions of temperature, speed, and payload. A

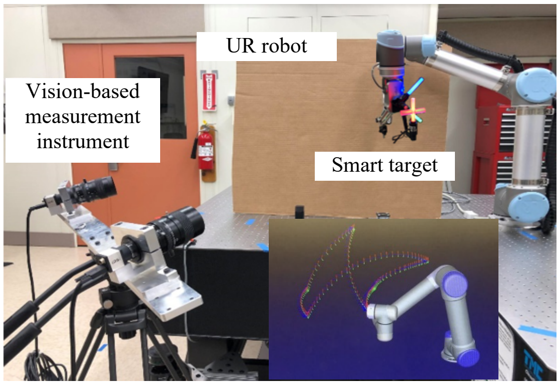


Figure 11. Use case setup

one-second motion halt was added to the program at the waypoint positions to observe how fast the robot can settle to stationary at different speeds. The smart target system measured the absolute positions of the robot arm. The absolute measurements from the smart target system were used to calculate the deviations from the nominal positions (designed positions). Simultaneously, the low-level controller data was collected. When accuracy degradations were found, controller data (includes target joint positions, actual joint positions, target joint velocities, actual joint velocities, etc.) can be used to analyze the root cause of the changes for PHM remedy. A timestamp was saved to align the smart target measurement data and the low-level controller data.

Tests were performed with a combination of payload (half payload and full payload), speed (half speed/full speed), and temperate (cold start to 2-hour warmup,10 degrees Celsius changes in waist joint). The test data set was published in [17]. The dataset shows that temperature and speed have more influence on the robot's pose deviation compared with the payload influences for this testing robot. The higher operating temperature made the position deviation worse. Overshot and position fluctuations were observed for the position deviation from 80 µm to 180 µm when the robot stopped at waypoints. The fluctuation may impact the part quality if the robot was performing some precision operations such as material removal. The dynamic performance of the robot needs to be carefully monitored. This use case demonstrated the feasibility of using the smart target to perform the robot accuracy degradation assessment based on different payload, speed, and temperatures. The developed sensor and methodology provided manufacturers a tool to quickly detect accuracy problems when a robot work cell was reconfigured, environmental condition changed, or an important action is to be performed.

V. Summary

This paper presented the advanced sensing development at NIST to support robot accuracy assessment and improvement. The hardware and software design of the smart target enables the high accuracy, high speed, dynamic, and continuous measurement of the moving robot arm's 6-D information.

Other than robot accuracy assessment, the smart target can be used for other applications that require high accuracy position and orientation as well. The dynamic high accuracy 6-D measurements are important in applications including tracking a moving object precisely, registering multiple instruments, providing feedback to unplanned adaptive control algorithms, etc. Future efforts are underway to develop additional industrial use cases for applications that require high-precision motions.

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] T. Kubela, A. Pochyly, and V. Singule, "Assessment of industrial robots accuracy in relation to accuracy improvement in machining processes," in *2016 IEEE International Power Electronics and Motion Control Conference (PEMC)*, 25-28 Sept. 2016 2016, pp. 720-725, doi: 10.1109/EPEPEMC.2016.7752083.

[2] A. Joubair, L. F. Zhao, P. Bigras, and I. Bonev, "Absolute accuracy analysis and improvement of a hybrid 6-DOF medical robot," *Ind. Robot.,* Article vol. 42, no. 1, pp. 44-53, 2015, doi: 10.1108/ir-09-2014-0396.

[3] W. Zhu, B. Mei, G. Yan, and Y. Ke, "Measurement error analysis and accuracy enhancement of 2D vision system for robotic drilling," *Robotics and Computer-Integrated Manufacturing,* vol. 30, no. 2, pp. 160-171, 2014/04/01/ 2014, doi: https://doi.org/10.1016/j.rcim.2013.09.014.

[4] Bhatt PM, Rajendran P, McKay K, Gupta SK., "Context-dependent compensation scheme to reduce trajectory execution errors for industrial manipulators," *2019 International Conference on Robotics and Automation (ICRA),* 2019 May, pp. 5578-5584.

[5] J. H. Jang, S. H. Kim, and Y. K. Kwak, "Calibration of geometric and non-geometric errors of an industrial robot," *ROBOTICA,* vol. 19, pp. 311-321, 2001.

[6] B. C. Jiang, R. Duraisamy, G. Wiens, and J. T. Black, "Robot metrology using two kinds of measurement equipment," *Journal of Intelligent Manufacturing,* vol. 8, no. 2, pp. 137-146, 1997/03/01 1997, doi: 10.1023/A:1018508805175.

[7] S. He *et al.*, "Multiple location constraints based industrial robot kinematic parameter calibration and accuracy assessment," *The International Journal of Advanced Manufacturing Technology,* vol. 102, no. 5, pp. 1037-1050, 2019, doi: 10.1007/s00170-018-2948-z.

[8] H. Du, X. Chen, D. Zhou, G. Guo, and J. Xi, *Integrated fringe projection 3D scanning system for large-scale metrology based on laser tracker* (Applied Optics and Photonics China (AOPC2017)). SPIE, 2017.

[9] J. Li *et al.*, "Calibration of a multiple axes 3-D laser scanning system consisting of robot, portable laser scanner and turntable," *Optik,* vol. 122, no. 4, pp. 324-329, 2011/02/01/ 2011, doi: https://doi.org/10.1016/j.ijleo.2010.02.014.

[10] C. Lightcap, S. Hamner, T. Schmitz, and S. Banks, "Improved Positioning Accuracy of the PA10-6CE Robot with Geometric and Flexibility Calibration," *IEEE Transactions on Robotics,* vol. 24, no. 2, pp. 452-456, 2008, doi: 10.1109/TRO.2007.914003.

[11] A. Wiles, D. Thompson, and D. Frantz, *Accuracy assessment and interpretation for optical tracking systems* (Medical Imaging 2004). SPIE, 2004.

[12] G. Qiao and B. A. Weiss, "Advancing measurement science to assess monitoring, diagnostics, and prognostics for manufacturing robotics," *International Journal of Prognostics and Health Management,* vol. 013, no. 013, 2016.

[13] G. Qiao, C. Schlenoff, and B. A. Weiss, "Quick positional health assessment for industrial robot prognostics and health management (PHM)," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 29 May-3 June 2017 2017, pp. 1815-1820, doi: 10.1109/ICRA.2017.7989214.

[14] G. Qiao and B. A. Weiss, "Industrial Robot Accuracy Degradation Monitoring and Quick Health Assessment," *Journal of Manufacturing Science and Engineering,* vol. 141, no. 7, 2019, doi: 10.1115/1.4043649.

[15] T. Luhmann, C. Fraser, and H.-G. Maas, "Sensor modeling and camera calibration for close-range photogrammetry," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 115, pp. 37-46, 2016/05/01/ 2016, doi: https://doi.org/10.1016/j.isprsjprs.2015.10.006.

[16] Z. Zhang, "Determining the Epipolar Geometry and its Uncertainty: A Review," *International Journal of Computer Vision,* vol. 27, no. 2, pp. 161-195, 1998/03/01 1998, doi: 10.1023/A:1007941100561.

[17] G. Qiao, "Degradation measurement of robot arm position accuracy," 2019, doi: https://www.nist.gov/el/intelligent-systems-division-73500/degradation-measurement-robot-arm-position-accuracy.

Communication Technology Problems and Needs of Rural First Responders

Kerrianne Morrison
National Institute of Standards and Technology
kerrianne.morrison@nist.gov

Yee-Yin Choong
National Institute of Standards and Technology
yee-yin.choong@nist.gov

Shaneé Dawkins
National Institute of Standards and Technology
shanee.dawkins@nist.gov

Sandra Spickard Prettyman
Culture Catalyst, LLC
Sspretty50@icloud.com

ABSTRACT

Although new technology may benefit rural first responders to help them serve their communities, to date little is known about what communication technology problems rural first responders most need addressed and what future technology they desire. To explore the context of use and communication technology problems and needs of rural first responders, semi-structured interviews were conducted with 63 rural first responders across four disciplines: Communications (Comm) Center & 9-1-1 Services, Emergency Medical Services, Fire Service, and Law Enforcement. Using qualitative data analysis, interview data were sorted into problems and needs categories. Rural first responders' greatest problems were with reliable coverage/connectivity, interoperability, implementation/information technology (IT) infrastructure, and physical ergonomics. Rural first responders' greatest need for new technology was to address their current problems, but they were interested in new technology that leverages real-time technology and location tracking. Implications for researchers and developers of public safety communication technology are discussed.

Keywords

Communication technology, first responders, public safety, rural communities, usability.

INTRODUCTION

Rural Environments and Incident Response

First responders in public safety disciplines, namely Communications (Comm) Center & 9-1-1 Services (COMMS), Emergency Medical Services (EMS), Fire Service (FF), and Law Enforcement (LE) personnel, respond to emergency incidents to serve and protect their communities. Although these professions face many dangers and difficulties, first responders in rural communities encounter unique challenges by nature of the rural areas they serve. To better understand these challenges and how to mitigate them, rural areas in the United States (U.S.; e.g., Ricci et al., 2003; Tiesman et al., 2007) and in countries around the world (e.g., Aftyka et al., 2014; Birdsey et al., 2016; Hang et al., 2004; Jennings et al., 2006) have been a topic of research, with many studies focusing exclusively on rural emergency response (e.g., Gamache et al., 2007; O'Meara et al., 2002; Oliver and Meier, 2004; Ramsell et al., 2019; Reddy et al., 2009; Roberts et al., 2014).

A commonality in much of this work is finding that rural first responders are tasked to serve small communities that span wide landmasses. According to the U.S. Census Bureau's definition, rural areas comprise 97% of the U.S.'s landmass, but only 19.3% of the population (Ratcliffe et al., 2016; U.S. Census Bureau). Because rural first responders must often cover wide distances, some studies of rural areas have found longer ambulance response times in rural areas (Aftyka et al., 2014; Jennings et al., 2006).

Rural first responders also respond to incidents resulting from the unique terrain of the area. Some rural areas are

Buchanan, Kerrianne; Choong, Yee-Yin; Dawkins, Shanee; Spickard Prettyman, Sandra. "Communication Technology Problems and Needs of Rural First Responders." Presented at 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021). May 23, 2021 - May 26, 2021.

impacted by seasonal weather, experiencing high rates of sporting injuries (e.g., skiing) during certain seasons (e.g., winter) (Birdsey et al., 2016) and high rates of injuries during times of the year with more severe weather (e.g., monsoons, rain) (Hang et al., 2004). Although injury-hospitalization and death percentages are higher in rural than urban areas (Coben et al., 2009; Tiesman et al., 2007), rural areas are often served by rural first responders with small staffs that rely on volunteers or community workers who often have less experience and training (Gamache et al., 2007; Roberts et al., 2014).

Rural Barriers to Technology

Although these environmental features make incident response different for rural first responders relative to their urban and suburban counterparts, rural first responders face additional challenges in utilizing the proper equipment to respond to incidents. Communication technology such as radios, cell phones, and mobile data computers (MDC) are one of the most important tools first responders use in incident response, allowing them to obtain information about incidents and coordinate the appropriate response (Choong et al., 2018). Unfortunately, rural first responders face two primary barriers that prevent them from accessing and using communication technology.

First, rural areas tend to lack the infrastructure needed to implement the latest communication technology (Federal Communications Commission (FCC), 2020). This lack of infrastructure results in a lack of broadband access in many rural areas (FCC, 2020) and slow broadband speeds in some areas (Meinrath et al., 2019; Perrin, 2019) that may ultimately prevent rural first responders from accessing and using technology for incident response. Moreover, the costs for buying, installing, and maintaining broadband infrastructure are high in rural areas (Strover, 2001; Yankelevich et al., 2017), sometimes due to the impact of natural geographic barriers (e.g., mountains) and harsh weather conditions on equipment (Pötsch et al., 2016; Surana et al., 2008).

Second, some work suggests that people in rural areas are reticent to adopt new technology. Despite many rural areas gaining more access to broadband infrastructure, the urban-rural broadband adoption gap continues to persist (Dickes et al., 2010; Department of Commerce (DOC), 2010; Whitacre, 2008). Although demographic disparities between rural and urban areas are related to these lower adoption rates (Whitacre, 2008), LaRose et al. (2007) suggest that broadband adoption in rural areas is predicated on individuals' prior experience with, expected outcomes using, and self-efficacy for using the internet. Relatedly, work examining non-internet users found that their primary reason against adopting broadband in their homes was that they did not have any interest or need for broadband (DOC, 2010). Although this was the top reason for both rural and urban households, a larger share of rural households than urban endorsed this belief. These studies suggest that people in rural areas may not adopt technology because the benefits of adopting new technology are not made clear to them (Dickes et al., 2010; LaRose et al., 2007). Unfortunately, this may result in preventing rural first responders from utilizing tools that may help them during incident response.

Opportunities to Address Barriers

Fortunately, new legislation has created opportunities for mitigating these challenges by developing new technology specifically for first responders. The U.S. Middle Class Tax Relief and Job Creation Act of 2012 (Public Law 112-96, 2012) provided funding and dedicated broadband to establish the Nationwide Public Safety Broadband Network (NPSBN). While NPSBN development is in progress, this network will improve broadband access for first responders by supplementing land mobile radio (LMR) with Long-Term Evolution (LTE) solutions. Currently, the public safety research and development community has focused on developing new communication technology for first responders to operate with the new network. By improving broadband access and developing new communication technology, rural first responders can better share critical information during emergencies and disasters (Comfort et al., 2004) as well as use new capabilities such as those that improve location information (Weichelt et al., 2019) and assist with providing care to people in remote locations ahead of ambulance arrival (e.g., telehealth; Ricci et al., 2003).

Although the NPSBN is poised to help address rural first responders' need for broadband infrastructure, a solution is still needed to ensure that rural first responders will adopt new technology. Recent work has emphasized adoption as a critical piece of developing new technology for rural first responders and communities (Gasco-Hernandez et al., 2019; Weichelt et al., 2019). These studies and others (Choong et al., 2018) emphasize that although technology shows great promise to help first responders, unless technology is developed with the first responders' context and needs in mind, this technology may not be adopted by them. The concept of including users of technology in technology development is central to human factors research and user-centered design (International Organization for Standardization (ISO), 2019). By understanding the user, a developer can design technology with the users' needs in mind (Hackos and Redish, 1998). Ultimately, this improves the usability of a product, increasing its efficiency, effectiveness, and satisfaction to the user (ISO, 2019). Therefore, in order to improve adoption of new technology, rural first responders must be directly included in research. This will ensure that technology is developed specifically for their context of use and that their current needs are accounted for when developing new technology.

Relevant Research on Rural First Responders

To date, most work examining rural first responders has examined their unique context of use. Studies examining the context for rural emergency and health care workers have found that rural emergency responders rely on community workers and volunteers (Greene et al., 2019; Roberts et al., 2014), feel overburdened (Iversen et al., 2002; Oliver and Meier, 2004), have fewer resources and equipment (Greene et al., 2019; Oliver and Meier, 2004; Pilemalm, 2018), and serve wide, remote, and geographically diverse areas (Greene et al., 2019; Iversen et al., 2002; Oliver and Meier, 2004). However, fewer studies have investigated how rural first responders perceive, interact with, and use communication technology.

The work that has assessed rural first responders' perceptions and use of communication technology has focused broadly on emergency and health care professionals, including nurses, emergency department workers, and EMS personnel (O'Meara et al., 2002; Reddy et al., 2009) as well as community citizens, volunteers, and organizations (Pilemalm et al., 2013; Ramsell et al., 2019). These studies find that emergency, health care, and volunteer personnel are hindered by their communication devices due to the lack of interoperability between the numerous devices they use (O'Meara et al., 2002; Reddy et al., 2009) and connectivity problems (Reddy et al., 2009) from a lack of infrastructure (O'Meara et al., 2002; Pilemalm et al., 2013). Recently Ramsell et al. (2019) found semi-professional emergency responders and community volunteers value smartphone application usability and interoperability to support communication during incident response.

Gaps in Past Work

Although these studies provide important insights, they have two important gaps. First, the studies that have assessed rural first responders' perceptions and use of communication technology are largely specific to healthcare professionals generally and EMS personnel. It is unclear if these same problems transfer to other types of rural first responder disciplines, or if other disciplines have different problems with communication technology. Second, many of these studies examined limited types of technology, focusing largely on network coverage and mobile devices (e.g., smartphones) rather than on other communication technology more broadly such as radios, MDC, and body cameras. More work is required to identify useful functionalities beyond networks and mobile phones and instead assess needs broadly across communication technology for rural first responders.

Current Study

In the current study we addressed these gaps in prior work by studying the communication technology problems and needs of rural first responders across four disciplines (i.e., COMMS, EMS, FF, and LE). We also built off prior work (Greene et al., 2019; Iversen et al., 2002; Oliver and Meier, 2004) to understand rural first responders' context of use. Focusing on hearing the voices of rural first responders is important as historically rural perspectives have been left out of research about rural environments (Chambers, 1994). Insights from this study can help developers to identify what shortcomings in current technology need to be addressed as well as where to invest future resources in developing technology for rural first responders. By ensuring solutions that are tailored to work within the unique environments in which rural first responders operate, rural first responders may be more eager to adopt and use new technology in incident response.

METHOD

We conducted an exploratory sequential mixed methods study with two phases. In Phase 1, qualitative interviews were conducted to comprehensively explore communication technology experiences of first responders. Findings from Phase 1 were then used to design the Phase 2 quantitative survey instrument. In this way, the exploratory nature of Phase 1 led to a broader representation of first responders in Phase 2. This paper only focuses on data and analysis from Phase 1 and details methods and results using qualitative analysis methods. There were many advantages to using a qualitative approach in this first phase of research. First, this approach allowed us to explore and probe the specific problems and needs experienced by first responders to provide deep insights into the experiences and perspectives of those first responders who participated. Although qualitative methods are exploratory in nature, these methods can be applied rigorously for analysis (see Saldaña, 2013) and achieve validity (see Shenton, 2004). Second, it allowed us flexibility to examine the nuances between different first responder disciplines. Third, this approach emphasized engaging directly with participants through semi-structured interview techniques. This allowed for examination of first responders' top of mind priorities and dynamic perspectives and ensured that the voices of first responders were included in the research.

Recruitment and Sampling

Sixty-three rural first responders across four disciplines (COMMS (n=18), EMS (n=6), FF (n=19), and LE (n=20)) participated in the study. Purposeful and snowball sampling were used to recruit first responders. This sample was a subset of a larger effort to recruit a national sample of first responders (Choong et al., 2018; Dawkins et al., 2019; Greene et al., 2019). Five of the ten Federal Emergency Management Agency (FEMA) (2020) regions in the U.S. were represented in the sample.

Procedure

The research team scheduled 45-minute semi-structured one-on-one interview sessions at the first responders' place of work (e.g., fire station). However, in order to maximize the number of rural first responders interviewed, a subset of the interviews was conducted in small groups rather than one-on-one. This resulted in 48 interview sessions total with 63 total first responder participants. Participants were informed that they could withdraw at any time, skip any question as needed, and decline to be audio recorded. They completed demographic questions before the interview sessions. All data were collected anonymously. Recorded interviews were transcribed then de-identified and assigned an interview number. The National Institute of Standards and Technology (NIST) Research Protections Office reviewed the protocol for this project and determined it meets the criteria for "exempt human subjects research".

Instruments

Interview Instrument

An interview instrument was developed to guide the discussion during the semi-structured interviews. The interview instrument focused on two high-level areas: 1) understanding first responders' context of work, and 2) identifying first responders' perceptions of and experiences with technology. To understand context of work, the interview instrument included questions and follow-up probes related to job tasks and routines, relationships with people they work with or for, and characteristics of the environment they work in. Questions about technology focused on what technology they use, what problems they have encountered, and what technology they wish they had for their jobs. The interview instrument was developed iteratively through a process with a literature review, pilot interviews with first responders, and feedback from first responders and human factors subject matter experts (see Choong et al., 2018 for full methodological details for Phase 1).

Demographics Questions

Demographic characteristics (i.e., discipline, years of service, area, location, gender, and age) were collected to ensure interview data reflected the diversity of rural first responders. Additionally, we asked two questions related to technology experience and adoption (see Figure 1 and Figure 2) to better understand rural first responders' familiarity with technology. Participants could select as many options as were applicable to their own experiences.

Participant characteristics

The sample was made up of 18 COMMS participants, 19 FF participants, 20 LE participants, and 6 EMS participants. Table 1 displays the number of participants across rural first responder disciplines by gender, age, and total years of service. The sample was less representative of female first responders than male first responders, with female first responders comprising only 13 participants, though this is consistent with low proportions of female responders in FF and LE disciplines nationally (Crooke, 2013; Evarts and Stein, 2020). Relatedly, the larger number of females in our COMMS sample was consistent with gender demographics for the discipline nationally (U.S. Bureau of Labor Statistics, 2019). A majority of the sample was between 36 and 55 years old and had a wide range of total years of service.

Table 1. Frequencies of Demographic Characteristics by Rural First Responder Disciplines

		COMMS	EMS	FF	LE	Grand Total
Gender	Female	10	1	0	2	13

	Male	8	5	19	18	50
Age (Years)	18-25	1	1	3	2	7
	26-35	2	1	3	5	11
	36-45	5	2	6	4	17
	46-55	8	1	5	8	22
	56-65	2	1	1	0	4
	over 65	0	0	1	1	2
Total Years of Service	1-5	2	3	3	3	11
	6-10	3	0	4	3	10
	11-15	4	1	2	2	9
	16-20	1	1	2	3	7
	21-25	1	0	5	7	13
	26-30	3	0	2	2	7
	Over 30	3	1	1	0	5
	No response	1	0	0	0	1
Total	Total Participants	18	6	19	20	63

Figure 1 and Figure 2[1] display rural first responders' experiences with using and adopting technology. Although nearly 83% indicated they could do most or all things with technology with some assistance, 19% indicated they had limited knowledge or needed help with technology. In looking at experience adopting new technology, nearly 40% mentioned they let others work out the kinks. Although 28.6% said they follow technology trends, nearly 21% either adopt new technology when theirs has died or it becomes required.

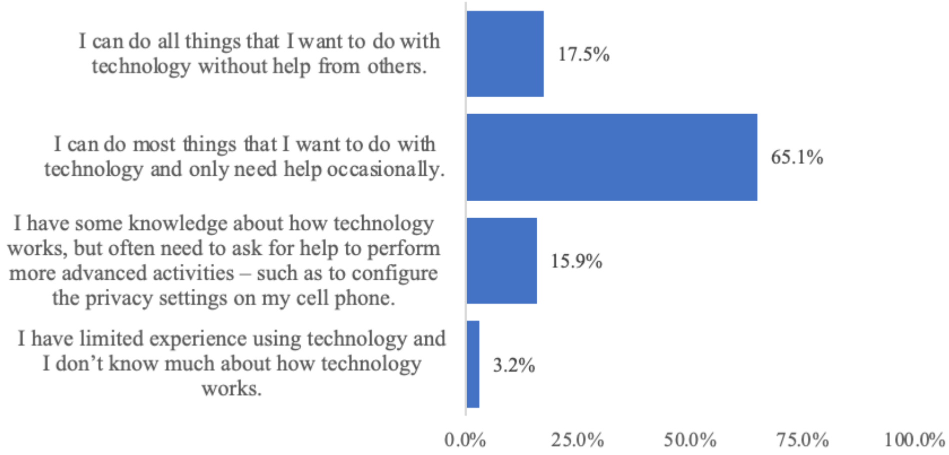


Figure 1. Experience with Technology.

[1] Because participants could select multiple responses and the total number of participants was used to compute percentages, responses sum to over 100%. One participant did not answer the questions.

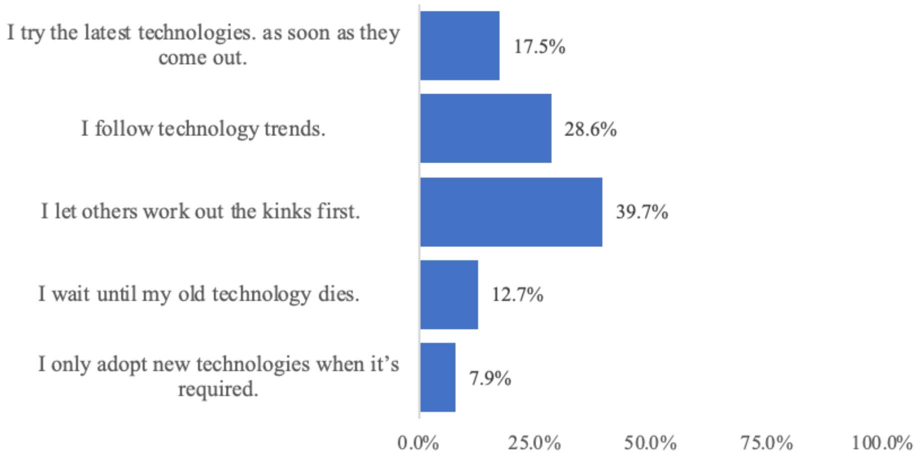


Figure 2. Experience Adopting Technology.

Qualitative Analysis

As part of the qualitative analysis process, transcripts were coded. Coding refers to assigning categories to participants' responses in order to reduce the data set so that it can be analyzed to find patterns and themes. The multidisciplinary research team first created an *a priori* coding list to be used for the initial coding of five randomly chosen transcripts from the entire project (see Choong et al., 2018). These five transcripts were independently coded by all team members, then the research team met to review their codes to determine if the codes were applied in consistent ways. This provided the opportunity to revise codes and operationalize how each should be applied, ultimately resulting in a finalized list of operationalized codes. The researchers coded all remaining transcripts using the final code list. The data associated with each code were extracted into separate files so that the relationships within and amongst the codes could be explored and themes identified.

The current study focused on the data from the transcripts of the 48 rural interviews with the 63 rural first responders, with codes related to: 1) communication technology problems and needs and 2) the context of use rural first responders operate within. First, to identify communication technology problems and needs, we reanalyzed responses initially coded into the "problem: technology" or "wish list" codes by further classifying responses into more specific categories and subcategories (see Dawkins et al., 2019). The "problems: technology" items were classified into the list of 18 technology problems displayed in Table 2. The same was done for the 15 "wish list" items displayed in Table 3. These categories and their corresponding subcategories were created for the larger research effort to identify the needs and requested functionalities that were most important to first responders (Dawkins et al., 2019). Here they were applied to the subset of the data with rural first responders. Two researchers independently identified the categories and subcategories for each response, with one researcher categorizing the problems and the other categorizing the needs. The research team then met to discuss, operationalize, and finalize the classifications. Second, to identify the rural context of use for problems and needs, we identified themes about the rural context from the extracted data (see Greene et al., 2019).

The qualitative results present themes using the direct quotes given by rural first responders. Quotes serve as exemplars and are representative of the data set as a whole. Each quote is followed by the reference to the participant in parentheses, including their discipline (i.e., COMMS, EMS, FF, or LE), area (R = Rural), and interview number (e.g., 001). Because participants were anonymous, identifiers are not tied back to a specific participant.

Buchanan, Kerrianne; Choong, Yee-Yin; Dawkins, Shanee; Spickard Prettyman, Sandra. "Communication Technology Problems and Needs of Rural First Responders." Presented at 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021). May 23, 2021 - May 26, 2021.

Table 2. Communication Technology Problems Categories and Subcategories

Category	Subcategories
9-1-1 Calls	Next Generation 9-1-1 (NG 911), caller location, nuisance calls
Audio Clarity	Hard to hear, audio feedback
Body Camera	Functional issues, physical issues
Connectivity	Reception, bandwidth issue
Disruption of Operations	Continuity of Operations (COOP), mobile operations
Implementation/Information Technology (IT) Infrastructure	Implementation/Installation issues, cost as a prohibitor, IT management, no user requirements collected/considered, public safety network reservations
Interoperability	External interoperability, internal interoperability
Microphone/Earpiece	Cord, earpiece, wireless microphones
Mobile Data Computer (MDC)/Mobile Data Terminal (MDT)	Navigation/mapping, functionality
Overwhelmed	Sensory overload, situational awareness
Physical Ergonomics	Robustness, battery problems, bulky and heavy, too many devices, physical discomfort, display size, safety concerns
Radio	Dead zones, traffic, channel switching, usability
Reliability	Unreliable technology, redundancy, unreliable transmissions
Security Constraints	Authentication, access control
Technology Outdated	Outdated, incomparable to personal technology
Technology Overrated	Problems with new technology, doesn't solve communication problems
User Interfaces	Ineffective and inefficient, alerting, modality
Video	Data issues, surveillance videos

Note. Categories and subcategories listed here are exhaustive but may not have been used across all disciplines. For example, "9-1-1 calls" only were coded for COMMS personnel.

Table 3. Communication Technology Wish List Categories and Subcategories

Category	Subcategories
All-In-One	Cell phones and/or radios, tablets, software and apps, general multifunctional devices, cameras
Communications Center Technology	Improved dispatch interface, multimedia data package, access to caller cell camera, large multi-view display
Functionality	Reliability, better coverage, clearer communication, improved functionality, longer battery life, faster devices
Futuristic	Media/Science-fiction influenced, smart buildings, face and object recognition software, self-driving vehicles, augmented reality (AR), emergency traffic light system
Integrated Gear/Wearables	Heads-up display (HUD), in-mask microphone/earpiece, responder vitals, personal protective equipment (PPE) technology
Interoperability	Software/hardware compatibility, interagency communication system, patient care report (PCR), body camera integration, interjurisdictional criminal data
Microphones/Earpieces	Wireless, specialized earpieces
Mobile Apps	Information references, discipline-specific apps
Physical Ergonomics	Smaller and lighter, fewer devices, robustness, larger devices
Radios	Channel switching, multiple talk groups, prevent accidental transmissions
Real-Time Technology	Live video and images - capture/live feed technology, traffic and navigation, drones, language translation, identification devices
Tracking	Responder location, caller location, search technology
Usable security	Single sign-on
User Interfaces	User friendly, hands free, non-verbal communication
Vehicles	Windshield HUD, built-in camera, automatic license plate reader, dashboard computer

Note. Categories and subcategories listed here are exhaustive but may not have been used across all disciplines. For example, the "body camera integration" subcategory in the "interoperability" category was only coded for law enforcement.

RESULTS

Technology Problems and Wish List Items Across Rural First Responder Disciplines

Technology problems are displayed in Figure 3 and wish list items in Figure 4. Across disciplines, rural first responders experienced the most problems with radios, reliability of devices, interoperability, implementation/IT infrastructure, and physical ergonomics. Their top wish list items were related to improving their current communication technology's functionality, specifically devices' reliability and coverage, as well as improving devices' interoperability and physical ergonomics. However, they also were interested in some futuristic technology, especially real-time technology and location tracking.

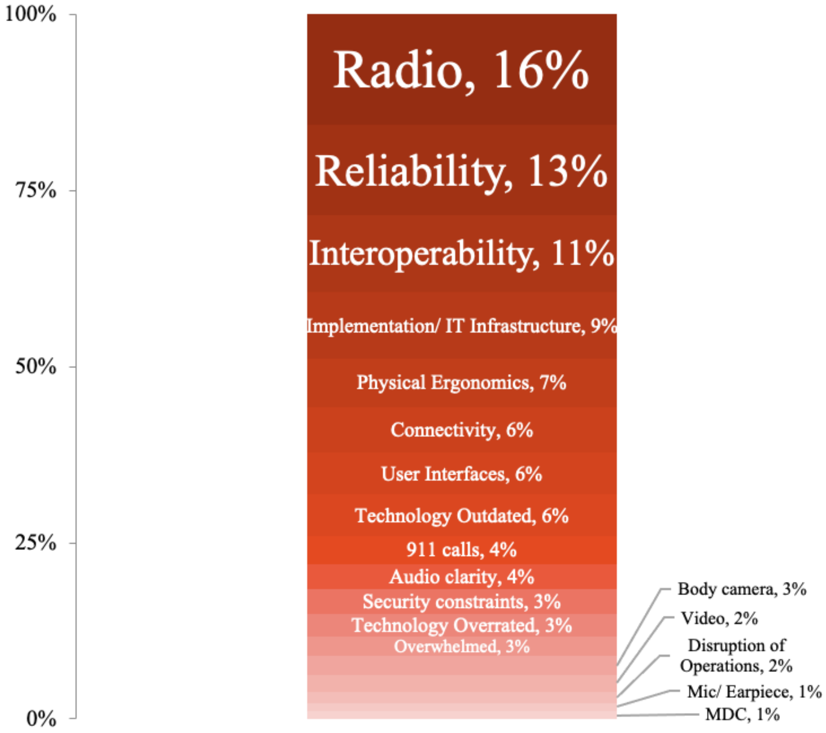


Figure 3. Technology problems across rural first responder disciplines.

Buchanan, Kerrianne; Choong, Yee-Yin; Dawkins, Shanee; Spickard Prettyman, Sandra. "Communication Technology Problems and Needs of Rural First Responders." Presented at 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021). May 23, 2021 - May 26, 2021.

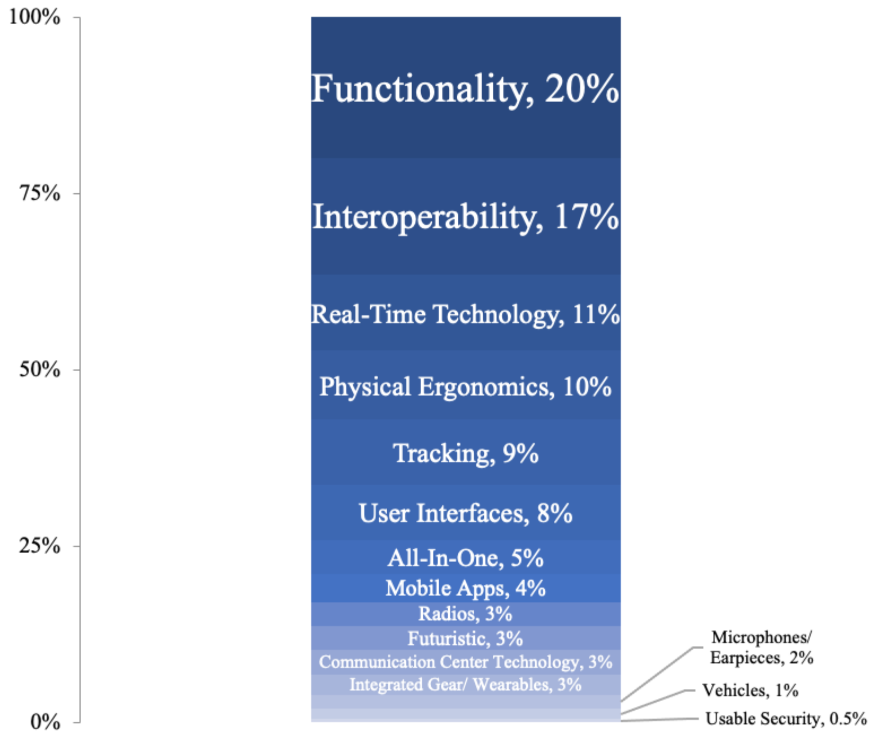


Figure 4. Wish list items across rural first responder disciplines.

Radio and Connectivity

Rural first responders experienced the most issues with dead zones preventing radio transmissions. Relatedly, rural first responders expressed problems with connectivity, especially in accessing bandwidth to gain reception for internet and cell phones. Some discussed dead zones in buildings or other structures, but many mentioned dead zones specific to rural terrain (e.g., mountains) that limit communication technology.

> Cell phones are great but again we're on a very rural county and if you get to the far ends of our county towards the east good luck. Radio traffic is null out there as well as cell phone is. (FF-R-049)

Rural first responders' most requested wish list item for improved functionality was improving radio and cell phone coverage in rural areas.

> ...You want your radios to work and you want your cell phones to work all over the county. I mean that's pretty much it…We did a missing person scenario down towards [county name redacted] in the national forest and nobody's radio worked...phone didn't work. (LE-R-048)

With access to wider coverage, rural first responders could improve the efficiency and effectiveness by which they communicate with their team members, transmit information to other responders and hospitals, and maintain a lifeline in dangerous situations.

Reliability

Many rural first responders felt their devices were unreliable, describing past experiences in which their communication technology did not work in the way intended.

> We have the [inaudible] MDTs [mobile data terminals], but I think we would call it a failed technology... We spend more time wasting time trying to keep that thing working than we do doing our job. So we've given up on it… (FF-R-019)

Finding a solution to reliability problems was also at the top of rural first responders' wish lists. They wanted to be able to trust the technology that they use, eliminating unnecessary distractions and stress.

Interoperability

Rural first responders described difficulties communicating among disciplines across rural areas and also during situations where they must work with other jurisdictions. The second most requested wish list item was improved interoperability, and many rural first responders were specifically interested in improved interagency communication. Improving communication interoperability for rural first responders with other disciplines, areas, and jurisdictions would help improve incident response, as well as increase information being shared with relevant parties.

> Speaker 2: I mean, I can't call [county name redacted], call on the cell phone. I can't call [another county name redacted]; we don't have their frequencies available, so it would all have to be relayed from us to here to County, to their dispatch to their officer and then back to the state again.
>
> Speaker 1: Which creates the delay you talked about earlier.
>
> Speaker 2: Right. And what is lost in translation. (LE-R-060)

Rural first responders also discussed that the numerous devices they use are not well integrated.

> I think my biggest gripes are that e-ticketing machine and just the fact that it's not well thought-out for the application. I don't think there's any reason why it couldn't be done on the phone that I already carry or the computer that's already in the car. (LE-R-018)

Rural first responders mentioned a need for more interoperability between software and hardware. Improving internal interoperability may decrease the amount of time to transmit information and may also reduce the burden, frustration, and confusion of using multiple devices.

Implementation/IT Infrastructure

Rural first responders described problems implementing and installing communication technology. One reason mentioned in the interviews was that many updates require access to the latest technology or use of broadband speeds to which many rural first responders do not yet have access. One COMMS worker described a situation where the communications center sought to use new technology but could not implement the new system because of the call center's outdated computer systems.

> So, you have to do an upgrade your phone system to make it compatible to go on NextGen…we are going to upgrade the phone switch, which we did. But then, our wall boards that tells us how many calls are waiting…they quit working. Well, we found out that the computers that are driving the wall boards are not powerful enough to drive the wall boards with the upgrade. (COMMS-R-019)

Many rural first responders discussed these issues with implementation as being related to a broader issue of funding.

> I mean funding is a huge issue…If a truck went down that truck's gone until we can save up the money or get a grant or figure out something to fix that truck. I mean we were living year to year as a department you know and that depended on the size of the department and the size of the town. (FF-R-048)

Because rural departments were often underfunded, cost may be a prohibitor for rural first responders in accessing, training for, updating, and replacing communication technology.

Buchanan, Kerrianne; Choong, Yee-Yin; Dawkins, Shanee; Spickard Prettyman, Sandra. "Communication Technology Problems and Needs of Rural First Responders." Presented at 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021). May 23, 2021 - May 26, 2021.

Physical Ergonomics

Physical ergonomics problems and needs captured a wide range of topics, with some related to rural first responders' number, size, and weight of devices, and others related to physical aspects of devices such as robustness, battery life, comfort, and safety concerns. Rural first responders discussed problems with devices' robustness in rural environments. Rural first responders must have robust equipment to meet the challenges of the incidents they respond to (e.g., extreme heat) and the environments they work within, as they often encounter difficult terrain such as mountains or rivers. For this reason, many rural first responders requested improved robustness as a wish list item.

> A lot of our computer sites are above 10,000, 11,000, 12,000 feet. You can't get to some of them unless it's summer. If one [of] them breaks down, well, just have to wait until the weather clears. I mean, we have ice loading on some of our tower sites such that it shuts them down… we wait until it warms up and it falls off…Sometimes we go up in a snow cab or the technician does. I mean, that's dangerous, expensive work… (EMS-R-008)

They also discussed having battery issues with their devices. Additionally, they requested having fewer devices to operate, as another problem was having too many devices. Although making devices more robust and integrated would be of interest to rural first responders across disciplines, some specific problems and needs differed depending on the discipline.

Technology Problems and Wish List Items for Each Rural First Responder Discipline

Although many of these problems and needs were common across all disciplines, each rural first responder discipline experienced unique problems specific to their job requirements and context of use. Each also identified specific wish list items that would be beneficial to their needs. The discipline-specific data presented below were emphasized within a discipline but were not unique to that discipline.

COMMS Responders

Rural COMMS personnel experienced unique problems by nature of the environment they work within. COMMS personnel do not respond on-scene; they instead take emergency calls and dispatch first responders to the scene. COMMS personnel discussed the difficulty in locating callers during 9-1-1 calls, as some rural areas do not have addresses. Unique terrains (e.g., mountains, rivers) also bring in seasonal tourists, causing an increase in visitors unfamiliar with the area who have difficulty identifying their location when calling 9-1-1. For these reasons, COMMS personnel saw benefits to using technology to improve caller tracking.

> I would want to have the ability to see where my callers are…And I wish I could have the technology to see like a straight path to guide my officers to my callers…There are times when we've had calls where people were in domestic situations and they couldn't really tell us everything so that would help a lot. (COMMS-R-014)

They were also interested in improving communications center technology, such as having improved dispatch interfaces, better multimedia data packages to receive information, and access to callers' cell phone cameras. Although they saw benefits to technology, COMMS personnel were wary of new technology. They could foresee negative impacts and new challenges that may come with new technology, especially Next-Generation 9-1-1, a digitally-based 9-1-1 system (see National 911 Program). COMMS personnel expressed concerns over both seeing graphic or inappropriate images in texts and needing to slow down their response time to communicate via text with callers.

EMS Responders

EMS personnel mentioned problems writing and sending patient reports to hospitals. They discussed their MDCs were often unreliable and would crash or otherwise fail to save a patient report. In addition to crashing, EMS personnel also discussed that sometimes when their computers were working, computers could still fail to send patient information due to disrupted connectivity.

EMS personnel discussed that reliable and usable technology was expensive, causing some departments to opt for outdated solutions. In some cases, EMS personnel discussed using pencil and paper for report writing rather than computers. When EMS personnel did have technology for writing and recording patient information, they were often frustrated by how difficult their systems were to use. In fact, EMS personnel sometimes spent more time writing a report than they needed to, and in some cases had to rewrite their reports. Moreover, some mentioned that because they did not have dedicated IT staff, they often wasted valuable time fixing their systems or finding alternate solutions.

> What happens when it doesn't work? What happens when we have trouble with it? Who fixes it? Because I can't just call downstairs to IT, okay? I've got a contractor that does our IT because we don't have an IT department. They're budgeted two days a week, maybe. (EMS-R-008)

For these reasons, EMS personnel expressed interest in reliable technology, as this improvement may save them time and frustration, allowing them to focus on their jobs.

FF Responders

FF personnel had difficulty with mics and radios during incident response. They had problems hearing their radios when there was external sound caused by the fire and alarms, and their mics picked up breathing and other sounds that made communications hard to hear. They also had physical ergonomics issues, ranging from problems with batteries to having devices that were not robust to rural incident response. Many also expressed that their technology was outdated.

> …When a fire is paged out here they may page out the appropriate response it may or may not go out over the radio. We have somewhat of an outdated underfunded antiquated communications here in our county. (FF-R-049)

Although improved functionality for current devices was the top requested wish list item, FF personnel also saw benefits to new technology that could provide them with live real-time information about rural fires. They also requested improved software/hardware compatibility that could link multiple devices to send information to a single source.

LE Responders

Use of body cameras is specific to LE personnel in their day-to-day work. Some expressed physical challenges securely attaching their body cameras to their uniform, as well as mentioned that they spend significant time and effort storing and uploading the cameras' information.

> [Regarding body camera videos] It can add quite a bit of time because for the most part the upload time is the real time…I think the longest recording I have was probably about 3 hours which it breaks it up into thirty minute intervals but it took almost 2 ½ or 3 hours for that one video to upload then I had 10 other ones that I had to upload so the upload speed is absolutely horrible. (LE-R-045)

LE personnel often mentioned challenges using devices that were bulky, too numerous, and not reliable due to battery problems. These ergonomics challenges were often specific to the equipment they use, such as e-ticketing devices and utility belt equipment.

Futuristic Functionalities

Beyond desiring technology to address their current problems, rural first responders most discussed wanting new technology that can provide real-time and location tracking information.

Real-Time Information

Rural first responders discussed that the ability to send and receive live video and images would be useful.

> Or that there's the ability that that camera would be tied to the MDC so that I could push a button, take a picture, and transmit that without sitting here and opening an email, figure out who's working today, who's going to get this email…(FF-R-008)

Accessing rich and detailed information in real-time would assist rural first responders with planning for and responding to incidents.

Access To Location Information

Many discussed that having tracking information would be useful for locating both 9-1-1 callers and first responders during incident response.

> And I think the mapping is a little -- like if we could somehow manage to afford, like live mapping or whatever…(COMMS-R-019)

Some rural first responders discussed that having information collected from previous incidents and locations as well as global positioning system (GPS) tracking of nearby responders and vehicles would improve preparing for and responding to incidents.

DISCUSSION

Rural first responders' primary communication technology problems were the lack of reliable coverage and connectivity, interoperability, implementation/IT infrastructure, and physical ergonomics of communication technology. This is consistent with work examining both rural (Greene et al., 2019; O'Meara et al., 2002; Pilemalm et al., 2013; Reddy et al., 2009) and urban and suburban public safety personnel (Dawkins et al., 2019). This suggests that research and development addressing these problems are likely to benefit all first responders. However, our findings suggest that the rural context of use must be considered in order to improve communication technology specifically for rural first responders.

Although urban and suburban first responders also experience dead zones and connectivity issues (Dawkins et al., 2019), the lack of broadband infrastructure and geographic dead zones are largely unique to rural areas. Most rural first responders in this study relied on using radios and cell phones to communicate, and when these devices were unable to connect, rural first responders had no way to coordinate with other responders in the area or acquire new information. Although broadband coverage has been improving (FCC, 2020), some areas that have coverage have slow speeds (Meinrath et al., 2019; Perrin, 2019). Developers should carefully consider the communication technology they develop for use in rural areas; until broadband access and speed is improved, some devices may not work as intended or at all. Therefore, researchers and designers should continue to consider how to increase coverage and connectivity of communication technology in rural areas.

Researchers and designers would also benefit rural first responders by developing devices that are robust to extreme weather and terrains and are developed with appropriate physical ergonomics requirements in mind for each first responder discipline. While FF personnel may need robust radios and mics adaptable to hot temperatures and loud scenes, LE personnel need body cameras and equipment for their belts that allow them to move easily. Moreover, rural first responders in this study had limited budgets that may preclude them from replacing technology often. Therefore, technology should also endure for a long period of time.

An important theme in the interviews was the additional burdens placed on rural first responders. Not only did they serve a wide area, but because of funding and resource allocations, they were often asked to do their jobs without proper equipment and personnel. Although technology has the potential to decrease these burdens by increasing the amount of information they have and decreasing time spent on tasks performed, we saw obstacles to this goal in our data. In many cases technology was an added burden, both mentally and physically to the day-to-day tasks of rural first responders. Some were distracted by managing and carrying many devices, and some spent extra time fixing technological problems and coordinating communication efforts between and across agencies. Therefore, designers should develop technology that will be easy to use and maintain, cost effective, lightweight, and integrated into technology they already possess. Additionally, in responding to the technology experience and adoption questions, approximately 61% of rural first responders indicated they would be hesitant to proactively adopt new technology. Developers must therefore ensure the benefits of new communication technology are made clear. By alleviating burdens caused by technology, rural first responders may more readily adopt new communication technology that will reduce their frustration and save them time. Such changes may help rural responders perform their jobs more efficiently, thereby decreasing the amount of time needed to respond to incidents. Because past research has found a relationship between response times and patient outcomes and satisfaction (Jennings et al., 2006; Persse et al., 2004; Rogers et al., 1999), improved efficiency may have benefits to rural first responders and their communities.

Taken together, it is unsurprising that when rural first responders were asked what new technology would benefit them, they wanted their current problems fixed rather than entirely new communication technology. However, this does not mean that rural first responders were uninterested in new or futuristic devices. Rather than seeing future technology as a way to improve communication, rural first responders saw more utility for technology to improve access to real-time information. These findings are consistent with prior work with urban and suburban first responders (Choong et al., 2018) and underscore the need for developers to address problems but also anticipate first responders' need for information.

It is important to note that these results were a part of the Phase 1 study, which was only the first phase of a larger sequential mixed methods study. Therefore, while the Phase 1 results were meant to explore the experiences of first responders and their context of work, the results were also intended for use in developing the Phase 2 survey. The qualitative results presented in this study were meant to deeply investigate a phenomenon by providing access into the world of those interviewed, in their own voices. Although the sample sizes for each discipline in this paper were small (e.g., n=6 for EMS), the data were rich and highly contextualized to the rural environment. Thus, the goal was not to generalize to the broader rural population, but to provide in-depth insights into the phenomenon from the perspective of those interviewed. Phase 2 was designed to expand upon the Phase 1 results with a larger sample. Phase 2 results can provide a broader representation of first responders and comprehensively capture communication technology usage, problems, and needs. Therefore, the ultimate aim of our research will be to use both exploratory findings from Phase 1 and broader representative findings from Phase 2 to inform the development and design of new technology for first responders. We encourage additional studies to continue to expand upon the communication technology experiences of first responders, especially those who work within rural environments. Specifically, work is needed to analyze differences in problems and needs depending on demographic subgroups (e.g., gender, age, volunteer status) as well as ensure findings can generalize to the broader rural population. Future work may also benefit from moving beyond self-report to using scenario-based assessments (see Pilemalm, 2018) to elucidate problems experienced during incident response, as well as testing new technology (e.g., real-time technology) with this population.

These limitations non-withstanding, results from this study highlight many ways communication technology can be designed and improved for rural first responders:

1. Better coverage and connectivity for cell phones and radios
2. Improved interoperability both for communicating across agencies and for integrating devices
3. Strong and long-lasting devices that work in extreme physical environments
4. Affordable devices that are easy to fix and inexpensive to train on

By continuing to study human factors in rural first responder populations, technology can be developed and improved for rural first responders. This could shift how rural first responders view, adopt, and use communication technology. Rural first responders may transition away from viewing communication technology as a problem and burden, and instead view communication technology as a trusted tool for more effectively and efficiently protecting and serving their communities.

REFERENCES

Aftyka, A., Rybojad, B., and Rudnicka-Drozak, E. (2014) Are there any differences in medical emergency team interventions between rural and urban areas? A single-centre cohort study, *Australian Journal of Rural Health,* 22, 5, 223-228.

Birdsey, M., Islam, M. R., and Barmare, A. (2016) Sporting injuries, seasonal trend and impact on rural Australian hospitals: Implications and recommendations, *Australian Journal of Rural Health,* 24, 6, 402-408.

Chambers, R. (1994) The origins and practice of participatory rural appraisal, *World Development,* 22, 7, 953-969.

Choong, Y., Dawkins, S., Furman, S., Greene, K. K., Prettyman, S. S., and Theofanos, M. F. (2018) Voices of First Responders – Identifying Public Safety Communication Problems, Findings from User-Centered Interviews, Phase 1, Volume 1. National Institute of Standards and Technology Interagency or Internal Report (NISTIR) 8216. doi: 10.6028/nist.Ir.8216

Coben, J. H., Tiesman, H. M., Bossarte, R. M., and Furbee, P. M. (2009) Rural–Urban Differences in Injury Hospitalizations in the U.S., 2004, *American Journal of Preventive Medicine,* 36, 1, 49-55.

Comfort, L. K., Ko, K., and Zagorecki, A. (2004) Coordination in Rapidly Evolving Disaster Response Systems: The Role of Information, *American Behavioral Scientist,* 48, 3, 295-313.

Crooke, C. (2013, July) Women in Law Enforcement, *Community Policing Dispatch*, Retrieved from https://cops.usdoj.gov/html/dispatch/07-2013/women_in_law_enforcement.asp

Dawkins, S., Choong, Y.-Y., Theofanos, M. F., Greene, K. K., Furman, S., Steves, M., and Prettyman, S. S. (2019) Voices of First Responders – Examining Public Safety Communication Problems and Requested Functionality, Phase 1 Volume 2.1. National Institute of Standards and Technology Interagency or Internal Report (NISTIR) 8245. doi: 10.6028/nist.Ir.8245

Dickes, L., A. , Lamie, D. R., and Whitacre, B. E. (2010) The Struggle for Broadband in Rural America, *Choices,* 25, 4.

Evarts, B., and Stein, G. P. (2020). National Fire Protection Association's (NFPA) US Fire Department Profile 2018. Retrieved from https://www.nfpa.org/News-and-Research/Data-research-and-tools/Emergency-Responders/US-fire-department-profile

Federal Communications Commission. (2020). 2020 Broadband Deployment Report. Retrieved from https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2020-broadband-deployment-report

Federal Emergency Management Agency (FEMA). (2020, September 15, 2020) Regions, Retrieved from https://www.fema.gov/about/organization/regions

Gamache, S., Hall, J. R., Ahrens, M., Penney, G., and Kirtley, E. (2007). Mitigation of the Rural Fire Problem: Strategies Based on Original Research and Adaptation of Existing Best Practices. Final Report of Cooperative Agreement EME-2004-CA-0187. Retrieved from https://www.semanticscholar.org/paper/Mitigation-of-the-Rural-Fire-Problem%3A-Strategies-on-Ahrens-Gamache/0199797c922c6d98ef5e7467c36ab24135800247?p2df

Gasco-Hernandez, M., Zheleva, M., Bogdanov, P., and Gil-Garcia, J. R. (2019) Towards a Socio-Technical Framework for Bridging the Digital Divide in Rural Emergency Preparedness and Response: Integrating User Adoption, Heterogeneous Wide-Area Networks, and Advanced Data Science*, Proceedings of the 20th Annual International Conference on Digital Government Research*, Dubai, United Arab Emirates.

Greene, K. K., Dawkins, S., Theofanos, M. F., Steves, M., Furman, S., Choong, Y.-Y., and Prettyman, S. S. (2019) Voices of First Responders – Examining Public Safety Communication from the Rural Perspective Phase 1, Volume 3. National Institute of Standards and Technology Interagency or Internal Report (NISTIR) 8277. doi: 10.6028/nist.Ir.8277

Hackos, J. T., and Redish, J. (1998) Chapter 2: Thinking about Users, In *User and task analysis for interface design* (pp. 23-50), Wiley, New York.

Hang, H. M., Byass, P., and Svanström, L. (2004) Incidence and seasonal variation of injury in rural Vietnam: a community-based survey, *Safety Science,* 42, 8, 691-701.

International Organization for Standardization. (2019). Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. In (Vol. ISO 9241-210).

Iversen, L., Farmer, J. C., and Hannaford, P. C. (2002) Workload pressures in rural general practice: a qualitative investigation, *Scandinavian Journal of Primary Health Care,* 20, 3, 139-144.

Jennings, P. A., Cameron, P., Walker, T., Bernard, S., and Smith, K. (2006) Out-of-hospital cardiac arrest in Victoria: rural and urban outcomes, *Medical Journal of Australia,* 185, 3, 135-139.

LaRose, R., Gregg, J. L., Strover, S., Straubhaar, J., and Carpenter, S. (2007) Closing the rural broadband gap: Promoting adoption of the Internet in rural America, *Telecommunications Policy,* 31, 6-7, 359-373.

Meinrath, S. D., Bonestroo, H., Bullen, G., Jansen, A., Mansour, S., Mitchell, C., . . . Thieme, N. (2019). Broadband Availability and Access in Rural Pennsylvania. Retrieved from https://www.rural.palegislature.us/broadband/Broadband_Availability_and_Access_in_Rural_Pennsylvania_2019_Report.pdf

Middle Class Tax Relief and Job Creation Act of 2012, Public Law 112–96, 126 Stat. 156. (2012, February 22, 2012) Retrieved from http://www.gpo.gov/fdsys/pkg/PLAW-112publ96/pdf/PLAW-112publ96.pdf

National Highway Traffic Safety Administration's Office of Emergency Medical Services National 911 Program. Next Generation 911, Retrieved from https://www.911.gov/issue_nextgeneration911.html

O'Meara, P., Burley, M., and Kelly, H. (2002) RURAL URGENT CARE MODELS: WHAT ARE THEY MADE OF?, *Australian Journal of Rural Health,* 10, 1, 45-50.

Oliver, W. M., and Meier, C. A. (2004) Stress in small town and rural law enforcement: Testing the assumptions, *American Journal of Criminal Justice,* 29, 1, 37-56.

Perrin, A. (2019, 8/18/2020) Digital gap between rural and nonrural America persists, *FactTank: News in the Numbers,* Retrieved from https://www.pewresearch.org/fact-tank/2019/05/31/digital-gap-between-rural-and-nonrural-america-persists/

Persse, D. E., Jarvis, J. L., Corpening, J., and Harris, B. (2004) Customer Satisfaction in a Large Urban Fire Department Emergency Medical Services System, *Academic Emergency Medicine,* 11, 1, 106-110.

Pilemalm, S. (2018) Participatory Design in Emerging Civic Engagement Initiatives in the New Public Sector: Applying PD Concepts in Resource-Scarce Organizations, *Association for Computing Machinery (ACM) Transactions on Computer-Human Interaction,* 25, 1, Article 5.

Pilemalm, S., Stenberg, R., and Andersson Granberg, T. (2013) Emergency Response in Rural Areas, *International Journal of Information Systems for Crisis Response and Management,* 5, 2, 19-31.

Pötsch, T., Schmitt, P., Chen, J., and Raghavan, B. (2016) Helping the Lone Operator in the Vast Frontier*,* *Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, Atlanta, GA.

Ramsell, E., Granberg, T. A., and Pilemalm, S. (2019) Identifying functions for smartphone based applications in volunteer emergency response*, Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2019)*, Valencia, Spain.

Ratcliffe, M., Burd, C., Holder, K., and Fields, A. (2016). Defining Rural at the U.S. Census Bureau. Retrieved from https://www.census.gov/library/publications/2016/acs/acsgeo-1.html

Reddy, M. C., Paul, S. A., Abraham, J., McNeese, M., DeFlitch, C., and Yen, J. (2009) Challenges to effective crisis management: using information and communication technologies to coordinate emergency medical services and emergency department teams, *International Journal of Medical Informatics,* 78, 4, 259-269.

Ricci, M. A., Caputo, M., Amour, J., Rogers, F. B., Sartorelli, K., Callas, P. W., and Malone, P. T. (2003) Telemedicine reduces discrepancies in rural trauma care, *Telemed J E Health,* 9, 1, 3-11.

Roberts, A., Nimegeer, A., Farmer, J., and Heaney, D. J. (2014) The experience of community first responders in co-producing rural health care: in the liminal gap between citizen and professional, *BMC Health Services Research,* 14, 1, 460.

Rogers, F. B., Shackford, S. R., Osler, T. M., Vane, D. W., and Davis, J. H. (1999) Rural trauma: the challenge for the next decade, *J Trauma,* 47, 4, 802-821.

Saldaña, J. (2013) *The coding manual for qualitative researchers* (2nd ed.), SAGE, Thousand Oaks, CA.

Shenton, A. K. (2004) Strategies for ensuring trustworthiness in qualitative research projects, *Education for Information,* 22, 63-75.

Strover, S. (2001) Rural internet connectivity, *Telecommunications Policy,* 25, 5, 331-347.

Surana, S., Patra, R., Nedevschi, S., Ramos, M., Subramanian, L., Ben-David, Y., and Brewer, E. (2008) Beyond pilots: keeping rural wireless networks alive*, Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*, San Francisco, California.

Tiesman, H., Zwerling, C., Peek-Asa, C., Sprince, N., and Cavanaugh, J. E. (2007) Non-fatal injuries among urban and rural residents: the National Health Interview Survey, 1997-2001, *Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention,* 13, 2, 115-119.

U.S. Bureau of Labor Statistics. (2019). HOUSEHOLD DATA ANNUAL AVERAGES 11: Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. Retrieved from https://www.bls.gov/cps/cpsaat11.htm

U.S. Census Bureau. Rural America, Retrieved from https://gis-portal.data.census.gov/arcgis/apps/MapSeries/index.html?appid=7a41374f6b03456e9d138cb014711e01

U.S. Department of Commerce Economics and Statistics Administration and the National Telecommunications and Information Administration. (2010). EXPLORING THE DIGITAL NATION: Home Broadband Internet Adoption in the United States. Retrieved from https://www.ntia.doc.gov/report/2010/exploring-digital-nation-home-broadband-internet-adoption-united-states

Weichelt, B., Heimonen, T., Pilz, M., Yoder, A., and Bendixsen, C. (2019) An Argument Against Cross-Platform Development: Lessons From an Augmented Reality App Prototype for Rural Emergency Responders, *JMIR Mhealth Uhealth,* 7, 3, e12207.

Whitacre, B. E. (2008) Factors influencing the temporal diffusion of broadband adoption: evidence from Oklahoma, *The Annals of Regional Science,* 42, 3, 661-679.

Yankelevich, A., Shapiro, M., and Dutton William, H. (2017) Reaching beyond the wire: challenges facing wireless for the last mile, *Digital Policy, Regulation and Governance,* 19, 3, 210-224.

NIST SP 1283
August 2022

Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques

Yaw S. Obeng, Papa K. Amoah
National Institute of Standards and Technology, Gaithersburg, MD 20899

Joe Smetana, Nokia
3201 Olympus Blvd, Irving, TX 75063

Richard Coyle; NOKIA-Bell Labs
600 Mountain Ave., New Providence, NJ 07974

Julie Silk; Morgan Allison, Keysight Technologies
1400 Fountaingrove Pkwy, Santa Rosa, CA 95403

Karl Sauter, Oracle Corporation
4120 Network Cir, Santa Clara, CA 95054

Tony Senese, Panasonic Electronic Materials Center
205 Ravendale Dr, Mountain View, CA, 94043

David Backen, TTM Technologies
850 Technology Way, Chippewa Falls, WI 54729

Robert Pennings, Flex
637 Gibraltar Court, Milpitas, CA. 95035

Bev Christian, High Density Packaging User Group International (HDPUG)
241 East River Road, St. George, Ontario, Canada N0E 1N0

[#]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.

Abstract

Historically, evaluations of solder joint failures and solder joint reliability have been done with direct current (DC) methods, using event detectors or data loggers for high-frequency circuits. Direct high-radio frequency (RF) measurements of signal paths are potentially more sensitive to incipient circuit (or solder joint) failure due to mechanical changes which may affect return loss, insertion loss, or phase angle, well before complete solder joint failure. Here are compared the fault detection capabilities and detection speeds, of direct current resistance (R_{DC}) to RF-based fault detection measurements to determine if RF signal loss could be a useful criterion for failure detection. In this paper, both high speed digital and analog RF circuits are considered.

Early S-parameter changes were observed, over time and thermal cycles, as the connectors were broken in from wear. Ultimately, the test circuits failed, due to cracks within the solder joints. The capacitance, and the capacitive reactance, of a partial crack in a solder joint was found to be substantially larger than the direct current resistance (R_{DC}) due to even a tiny remaining amount of intact solder joint. The low resistance so dominates the circuit that the circuit changes are unmeasurable

This paper was first presented at the 2021 IPC Apex Expo Technical Conference and published in the 2020 Technical Conference Proceedings.

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

by RF techniques until the crack is fully open. Thus, while the failures in high-frequency circuits from solder joint cracking are expected to occur simultaneously, or even after the DC failures occur, they are undetectable until total decohesion within the solder joint. As a result, the detection of failures using RF monitoring (S-parameters) lags that of failure detection by DC resistance measurement when evaluated by cycles to failure.

The results presented in this paper should be of benefit to component manufacturers working to determine the reliability of their components on test boards, their original equipment manufacturer (OEM) customers concerned about the components and their attachment to actual product circuit boards, and EMS and test labs providing services to component suppliers and OEMs.

Introduction

Many transmission lines on a high-speed server or network board have signal loss budgets of 10dB or less. A 1- or 2- dB loss on a solder joint could result in signal integrity failures. This is comparable to the loss typical in a well-designed FCBGA package. Traditional reliability measures such as DC-resistance (R_{DC}) may not adequately capture signal loss due to solder joint failures. Thus, new approaches to metrology need to be investigated. Kwon et al have demonstrated the application of microwave measurements in detecting incipient pre-catastrophic solder joint failure fractures in actual solder joints (as opposed to internal component electrical connectivity)[1-5]. High-frequency RF measurements of signal paths are potentially more sensitive to incipient circuit (or solder joint) failure[1-10]. Elsewhere, is researchers have correlation of DC-resistance due to device and material failures with changes in the microwave propagation characteristics[8].

Low-frequency measurements are done by direct current resistance (R_{DC}) measurements most often using event detectors or resistance data loggers. Low frequency resistance (R_{DC}) measurements are insensitive to incipient failures in emerging interconnects because of the large volume via fill in features such as through substrate vias (TSVs). Illustratively, void formation in TSVs was difficult to measure with electrical resistance change, while the onset of void formation results in impedance changes that are easily measured with the insertion losses of broadband microwave spectrum[9-10]. Furthermore, the phase changes in the propagating microwave can yield additional mechanistic information, such as changes in dielectric properties of the materials of construction, if they occur[6]. Prior to this test, it was anticipated that failures in RF circuits may occur because of mechanical changes that affect return loss, insertion loss or phase angle, well before complete solder joint failure.

RF testing may be done by monitoring scattering parameters (S-parameters) with a vector network analyzer (VNA) or high-speed time domain reflectometry (TDR) measurements. A TDR can measure the impedance changes and detect impedance discontinuities in the time domain, similar to what S-Parameters capture in the frequency domain[2-3]. These methods can measure travel time between source and defect site, attenuation constant (a measure of the total microwave energy loss from the dielectric and skin effect losses), RF signal phase, and group delay changes related to the changes in the dielectric properties of the device under test. Table 1 lists some of the relationships between common electrical measurements and the S-parameters commonly obtained in RF measurements[12]. These electrical quantities can be further transformed to provide more mechanistically relevant metrics. These can include Insertion Loss (S21) or Return Loss (S12).

Figure 1 illustrates some of the many chemical and mechanical changes that can result in changes in the microwave propagation characteristics in prototypical I/O circuits from previous published work[8]. Conceptually, the pre-catastrophic failure may result in a change in impedance of the device under test (DUT). Thus, the impedance can be used as a monitor of the solder joint reliability. The impact of an incipient mechanical crack will be discussed in detail below on a microwave bridge circuit.

This paper was first presented at the 2021 IPC Apex Expo Technical Conference and published in the 2020 Technical Conference Proceedings.

Table 1 Relationship between S-parameters from microwave measurements and some common electrical parameters

$$R=\mathrm{Real}\left(Z_0\frac{(1+S_{11})(1+S_{22})-S_{12}S_{21}}{2S_{21}}\right)$$

$$L=\frac{\mathrm{Imag}}{\omega}\left(Z_0\frac{(1+S_{11})(1+S_{22})-S_{12}S_{21}}{2S_{21}}\right)$$

$$G=\mathrm{Real}\left(\frac{(1+S_{11})(1-S_{22})+S_{12}S_{21}-2S_{21}}{Z_0((1+S_{11})(1+S_{22})-S_{12}S_{21})}\right)$$

$$C=\frac{\mathrm{Imag}}{\omega}\left(\frac{(1+S_{11})(1-S_{22})+S_{12}S_{21}-2S_{21}}{Z_0((1+S_{11})(1+S_{22})-S_{12}S_{21})}\right)$$

This paper directly compares the fault detection capabilities and detection speeds, of R_{DC} to RF, and determines if RF measurement of signal loss at different frequencies is a useful criterion for failure. Here, the attempt was made to use these microwave-based techniques on a customized PCB to study solder joint reliability, and to compare DC measurements to RF measurements. In this work, solder joints were stressed using accelerated thermal cycling, and changes in the microwave propagation characteristics of the circuits (usually represented as S-parameters) were measured. The changes in the S-parameters were leveraged to study the thermo-mechanical reliability associated with the thermal cycling of the purposely designed printed circuit boards.

The benefits were expected to be, but not limited to:

- Statistically determine the differences between DC and RF performance of solder joints
- Perhaps change our definition of "failed" solder joint
- Potentially explain many No Failure Found (NFF) field returns.

Materials

Printed Circuit Board (PCB) Design

Figure 2 shows the conceptual routing of the RF traces on the PCB design developed for this project. The figure shows the RF net (only), with a through-hole SMA connector routing through internal traces. The PCB boards were connected to a VNA with 40 GHz-rated semi-rigid RF cables through connectors rated for 27 GHz. Practical Components part number WLP256-.5-8MM-DC-SAC305. WSCSP dies were used as device components. All circuit packs consisting of the boards and the two components per board were built with SAC305, Indium 8.9HF1.

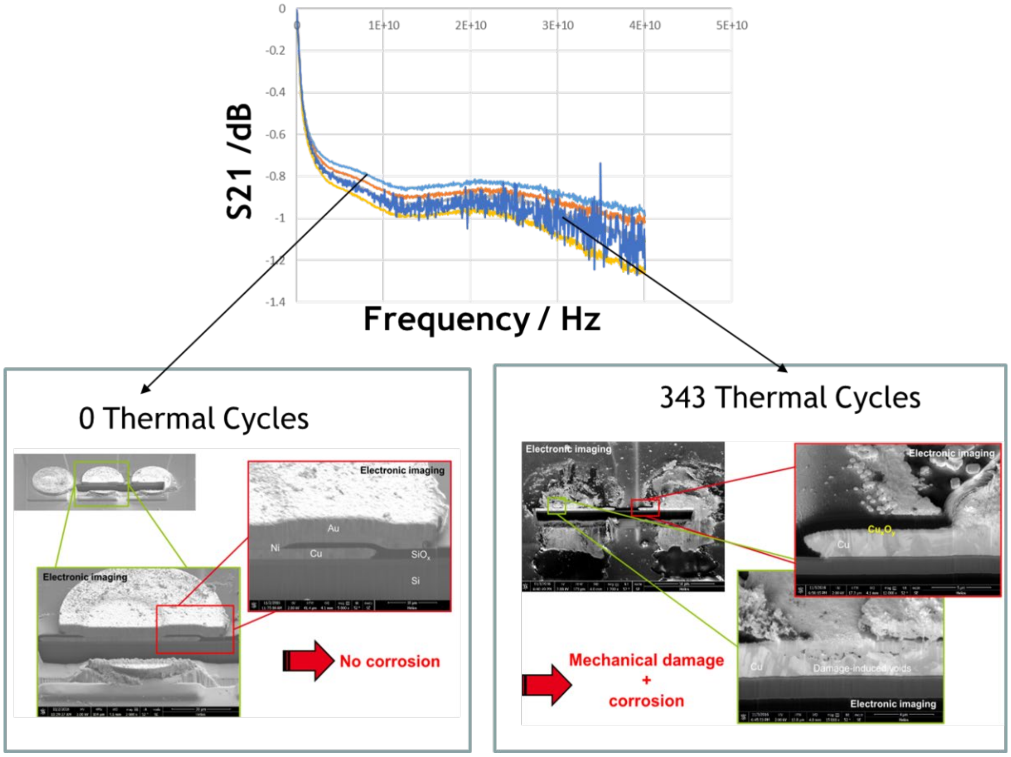


Figure 1: Illustration of the impact of thermal cycling on S-parameters of copper interconnects in open air due to corrosion (taken from Reference 5)

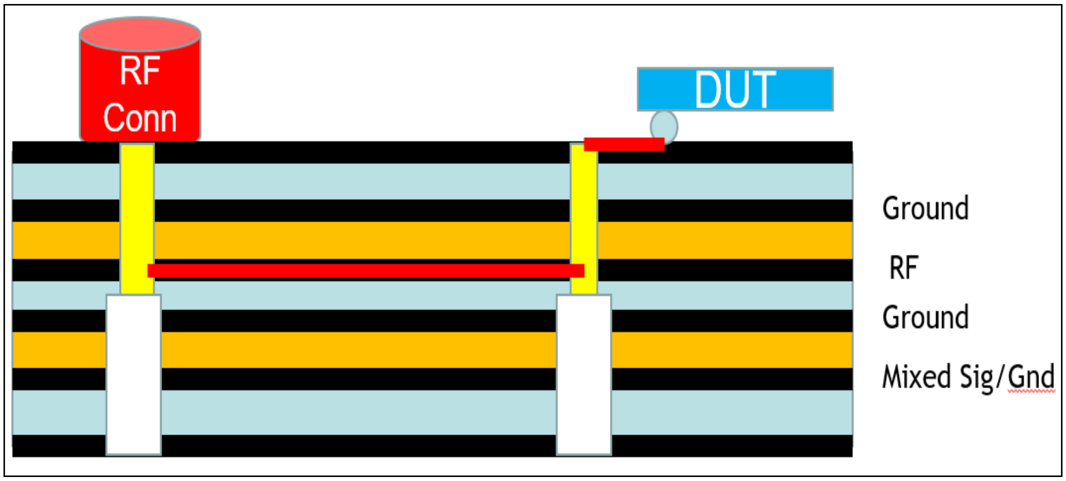


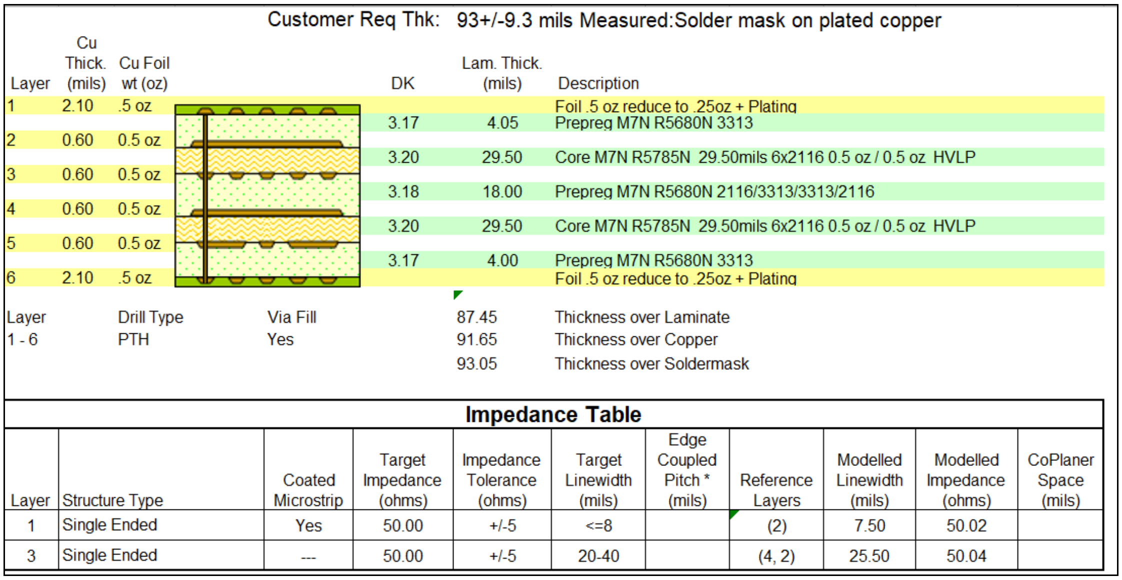
Figure 2: Conceptual routing of the RF traces

The top layer artwork design and the actual front and back of the completed PCB are shown in **Figure 2** while **Figure 3** shows the detailed PCB stack-up including trace width details. **Figure 4** shows the outer top surface of the test boards to illustrate the placement of the various test devices. The test board had the following features:

- Size: 16.5 cm X 17.8 cm, and 0.24 cm thick (i.e., 6.5” x 7”x 4, 93 mil thick)
- A single laminate material: Panasonic Megtron 7N (R5675N core, R5680N prepreg) with a board stackup as shown in Figure 3 was used.

This paper was first presented at the 2021 IPC Apex Expo Technical Conference and published in the 2020 Technical Conference Proceedings.

- Immersion silver finish
- There are 2 identical components per board. One wired for DC resistance measurements, and one for S-Parameter measurements.
- For the 4 DC corner circuits, wires were soldered to the appropriate PTHs.
- For the 4 RF corner circuits Molex SMA Jacks, FLANGE with 0-80 THD, 50 OHM 27 GHZ EWR-3690 SMA-J/R/F connectors were attached to the boards.



Customer Req Thk: 93+/-9.3 mils Measured:Solder mask on plated copper

Layer	Cu Thick. (mils)	Cu Foil wt (oz)	DK	Lam. Thick. (mils)	Description
1	2.10	.5 oz			Foil .5 oz reduce to .25oz + Plating
			3.17	4.05	Prepreg M7N R5680N 3313
2	0.60	0.5 oz			
			3.20	29.50	Core M7N R5785N 29.50mils 6x2116 0.5 oz / 0.5 oz HVLP
3	0.60	0.5 oz			
			3.18	18.00	Prepreg M7N R5680N 2116/3313/3313/2116
4	0.60	0.5 oz			
			3.20	29.50	Core M7N R5785N 29.50mils 6x2116 0.5 oz / 0.5 oz HVLP
5	0.60	0.5 oz			
			3.17	4.00	Prepreg M7N R5680N 3313
6	2.10	.5 oz			Foil .5 oz reduce to .25oz + Plating

Layer	Drill Type	Via Fill		
1 - 6	PTH	Yes	87.45	Thickness over Laminate
			91.65	Thickness over Copper
			93.05	Thickness over Soldermask

Impedance Table

Layer	Structure Type	Coated Microstrip	Target Impedance (ohms)	Impedance Tolerance (ohms)	Target Linewidth (mils)	Edge Coupled Pitch * (mils)	Reference Layers	Modelled Linewidth (mils)	Modelled Impedance (ohms)	CoPlaner Space (mils)
1	Single Ended	Yes	50.00	+/-5	<=8		(2)	7.50	50.02	
3	Single Ended	---	50.00	+/-5	20-40		(4, 2)	25.50	50.04	

Figure 3: Detailed PCB stackup including trace width details

Test Method

The daisy-chained components and the test circuit boards enabled electrical continuity testing after surface mount assembly and in situ, continuous monitoring during thermal cycling by DC measurements for one component and by RF measurement for the other component on each board. Thermal cycling was done in accordance with the IPC-9701A guidelines. The solder joints were monitored by DC means using the following criteria: a data logger set at a resistance limit of 3 ± 0.2 Ω. The RF measurements were obtained with a Keysight vector network analyzer (VNA). The switching system for the RF measurements is shown in **Figure 5**. The failure data are reported as characteristic lifetime eta η (the number of cycles to achieve 63.2% failure) and slope β from a two-parameter (2-P) Weibull analysis.

The temperature cycling profiles used in this investigation was 0 to 100°C testing with 915 cycles completed. This profile was selected to address the requirement of a specific industry or market segment as defined in standard IPC-9701A, specifically, telecom, represented by technical committee 1 (TC1). The temperature ramp rate was nominally 10 ºC/minute.

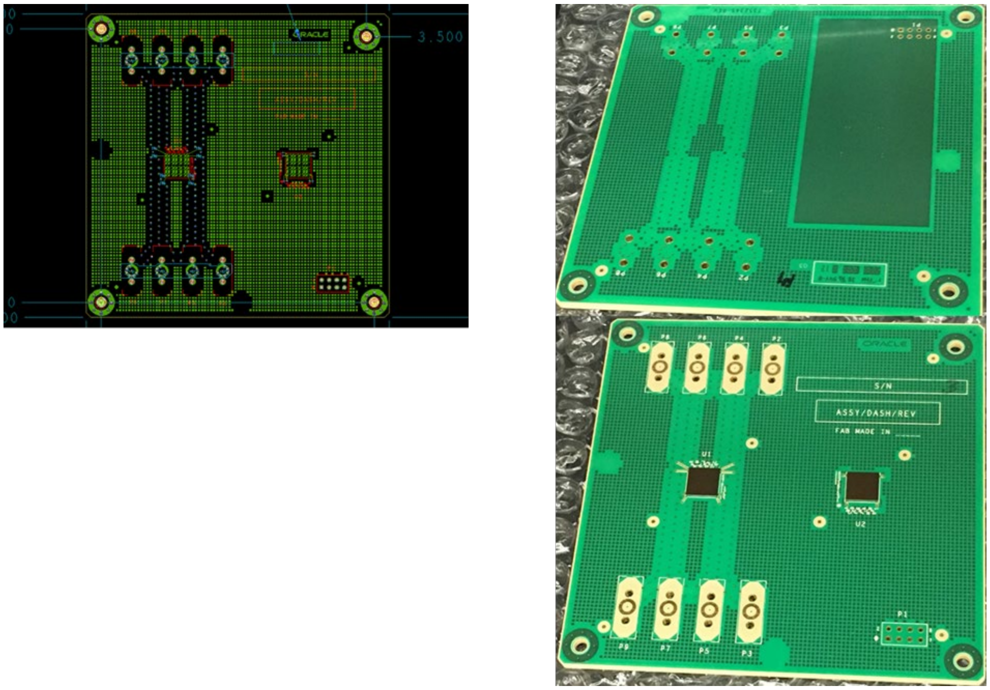


Figure 4: Outer Top Surface of the RF Failure Detection Test Board

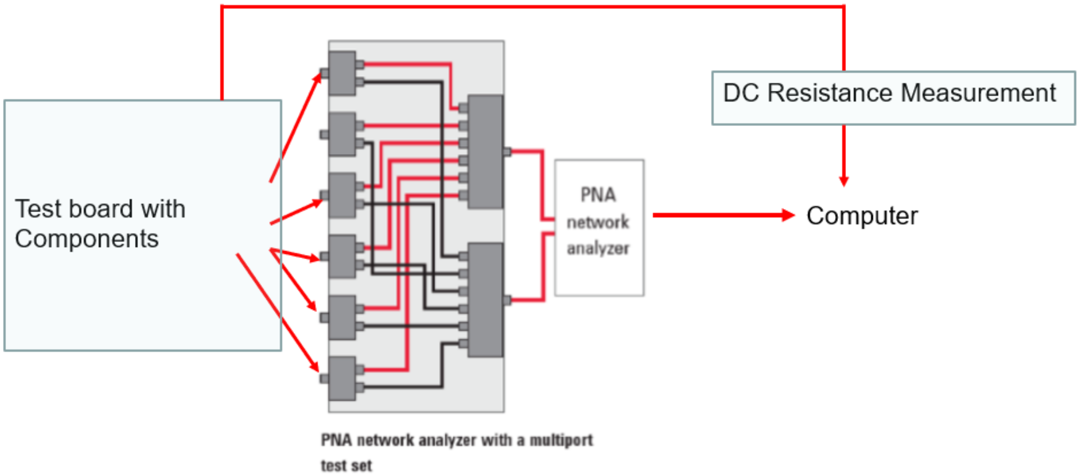

Test board with Components
DC Resistance Measurement
PNA network analyzer
Computer
PNA network analyzer with a multiport test set

Figure 5: The switching system for the RF measurements

Results and Discussion

Failure Definition

DC-Resistance Monitoring: The direct current (i.e., low frequency) electrical resistance changes in the test boards were monitored at the corners of a dedicated circuitry (U2) mounted on the test board. As shown in **Figure 6** below, the DC resistance (R_{DC}) of all the four corners of the DC- test die remained constant for about the first 700 thermal cycles, and then changed rapidly as the joints failed with increasing thermal cycling. The onset of rapid R_{DC} increase was used as the failure point as shown in **Figure 6**.

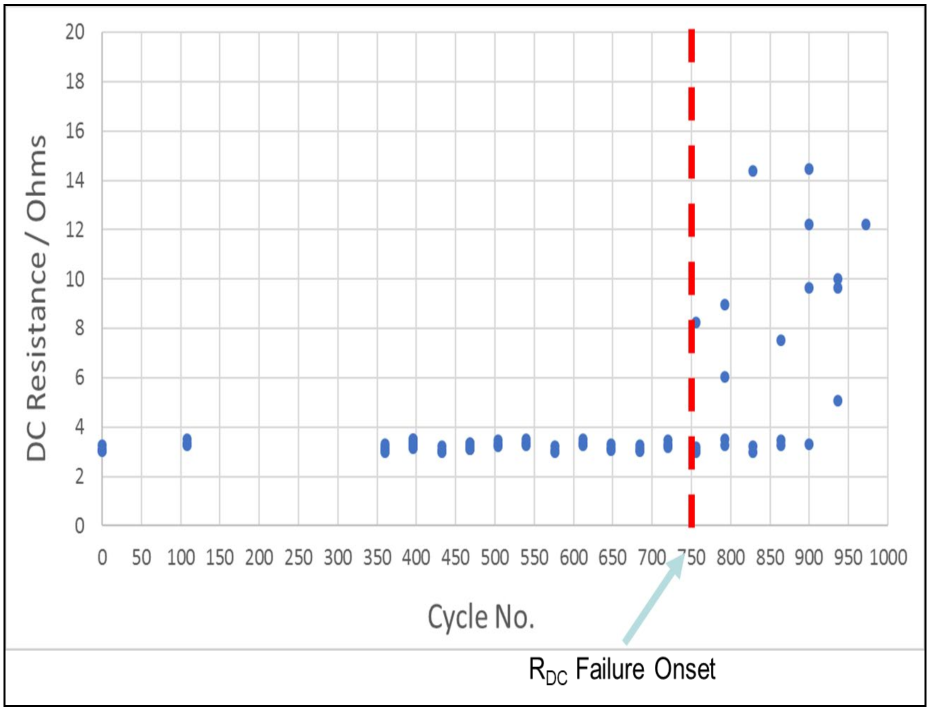


Figure 6: The evolution of the DC resistance of all the monitored DC-channels on a typical test board as a function of the number of thermal cycles.

Analysis of RF Data

Inspection of the accumulated experimental data revealed that both return loss, S11, amplitude and unwrapped phase (UP) angle afforded identical failure onset time at both frequencies. The following analyses focused only on the unwrapped phase (UP) angle of the return loss (S11) at 1 GHz for all the RF channels, **Figure 7** shows the evolution of the microwave signal phase at 1GHz for a single corner of a typical DUT as a function of thermal cycling. The figure shows several discrete breaks in the data evolution. The data for the first 300 cycles was used to characterize the break-in of the RF switches from mechanical wear of the electromechanical contacts.[13]. The initial 300 data points were used to set control limits of the 'aged' circuit and the onset of deviation from these control limits was used as the failure point. The stress duration before failure, using these definitions of failure as defined in Figures 6 and 7 respectively were used to generate the Weibull distribution plot in **Figure 8**.

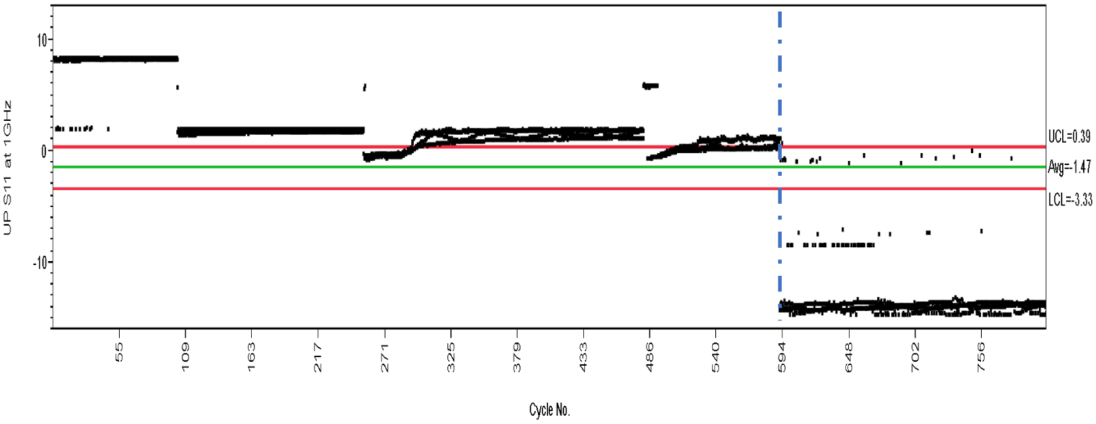


Figure 7: RF Monitoring (using S11 Phase changes at 1 GHz)

This paper was first presented at the 2021 IPC Apex Expo Technical Conference and published in the 2020 Technical Conference Proceedings.

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

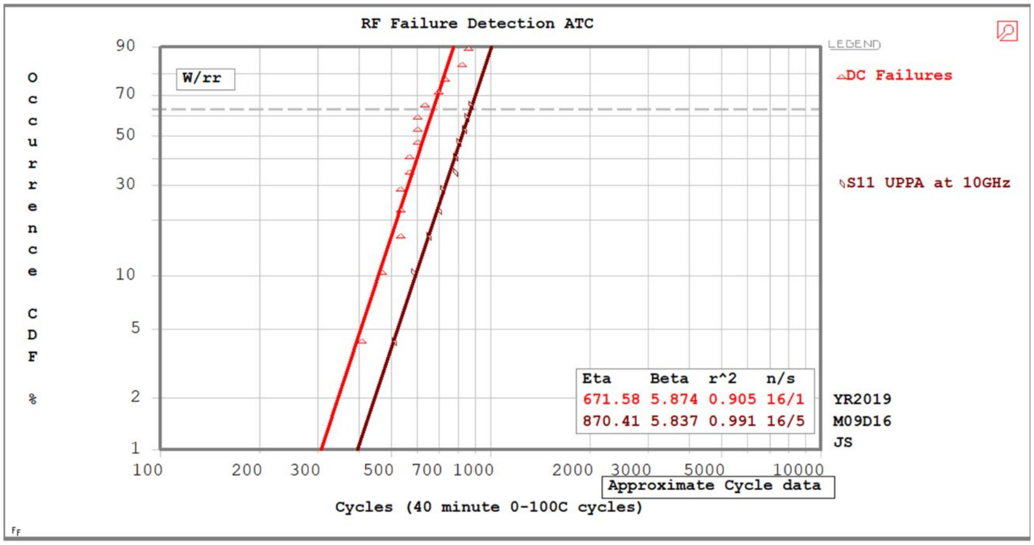


Figure 8: A Weibull distribution of the device failure times for R_{DC} and RF monitored components from Table 4. Clearly, the DC failures proceeded the RF failures.

Figure 9 shows the cross-section failure analysis of the failed waferscale CSP nets on a single board. These failed in all cases by solder fatigue as expected. Additional “Dye and Pry” analyses showed that both the RF and DC monitored components had similar failure patterns.

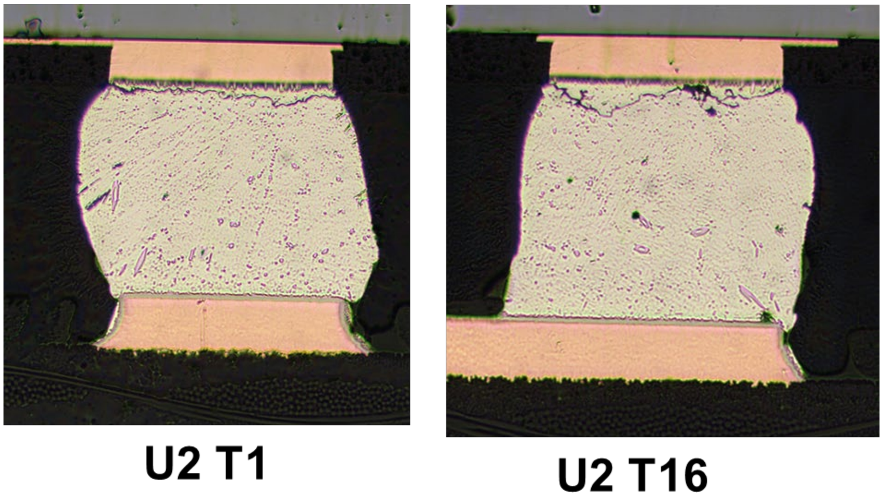


Figure 9: Cross-section of failed solder joints from two sample channels on a single test board. Similar for both the RF and DC monitored nets

Discussion of results and relevant electrical parameters

Previously published work on this subject reported that high frequency failures of the solder joint precede the DC failures[1-3]. Although the previous work was based on a very limited sample size, similar results were expected. In contrast, in the current report, the RF failures occurred after the DC failures. Current observations are explained at this point. When a crack opens or begins to open in a solder joint, it forms an air dielectric ($\mathcal{E}_r \approx 1$) capacitor, very similar to a parallel plate capacitor. The appearance of a change in a microwave measurement requires a change in the DC resistance (R_{DC}) of the joint. As the crack propagates, R_{DC} increases, and so does the capacitance (C). The prominence of microwave vs DC failure relates to the relative rate of increase of R_{DC} and C change. If R_{DC} (DC resistance of the crack) stays small until the crack fully propagates across the joint, then it will not appear in a microwave measurement. As shown by the following theoretical analysis, if R_{DC} increases before the crack propagates all the way, then it would be observed in a microwave analysis and as well as in a careful DC measurement.

The value of the parallel plate capacitor is per the following formula:

$$C = \varepsilon_r \varepsilon_0 \, A/d \qquad \text{(Equation 1)}$$

Where C = Capacitance (in Farads)
ε_r = relative permittivity (dielectric constant)
$\varepsilon_0 = 8.85 \times 10^{-12}$ F/m (permittivity of free space)
A = Area of the capacitor plates (in meter^2)
d = distance between the plates (in meters)

The capacitive reactance of this capacitor, in ohms, varies as per the following formula:

$$X_c = 1/(2\pi fC) \qquad \text{(Equation 2)}$$

Where X_c is the Capacitive reactance in ohms, π is a constant (3.1416), f is frequency (hertz), and C is Capacitive reactance in ohms

From Equation 2, the capacitive reactance is smaller at the higher the frequencies, such that the capacitor acts like a short circuit at high frequencies. It varies per the curve shown below in Figure 10, below.

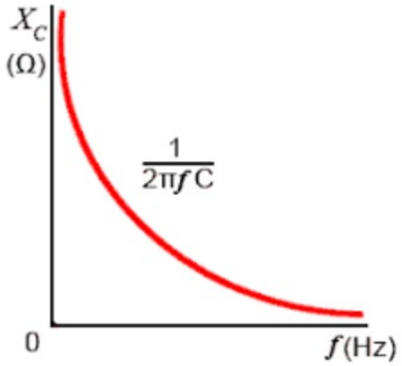


Figure 10: Capacitive reactance as a function of the frequency

It is obvious from equation 2 that the capacitive reactance also varies inversely to the capacitance, viz., the larger the capacitance, the smaller the capacitive reactance. Equation 1 shows that the capacitance value increases with decreasing separation distance between the plates of the capacitor. A typical fatigue solder joint crack has a very small distance between the "plates" resulting in a relatively large capacitance, even with the very small capacitor plates of a solder joint, until the solder joint fully opens and the joint separates. The crack surfaces are very nearly in contact up until the joint fully fails and the 'plates' become physically detached. The crack dimension, typically less than 1μm – on the order of about 0.1 μm (0.1 x 10^{-6} m) -- can be measured from the scanning electron micrographs of failed solder joints.

For the 0.5 mm waferscale CSP used in this testing, the crack area, which would be the capacitor plate size, is going to be very close to that of the pad size (0.3mm diameter), probably larger since the crack path is not straight. This is 7.069 x10^{-8} m^2. The rough approximate resulting capacitance for a fully open solder joint is

$$((8.85 \times 10^{-12}) \text{ x } (7.069 \text{ x}10^{-8}))/\ 0.1 \text{ x } 10^{-6} = 6.26 \text{ x } 10^{-12} \text{ Farads.}$$

The capacitive reactance for this at 10GHz is $1/(2\pi \text{ x } (1\text{x}10^{10}) \text{ x } (6.26 \text{ x } 10^{-12})) = 2.5$ ohms.

For a non-fully open solder joint, and, also for an "open" solder joint (the resistance of the "open" solder joint is considerably higher – usually over 1000 ohms), looking specifically at the crack and the crack alone (ignoring the rest of the circuit), the equivalent circuit schematic is shown in **Figure 11** below.

Here R is the resistance of the crack, L is the inductance of the crack, and C is the capacitance of the crack.

Considering the geometry of the solder joint crack, it is safe to consider the inductance of the crack to be zero. The capacitance of the crack was roughly estimated above. Until the crack is open, the resistance of the crack is going to be extremely low. The resistivity of tin is approximately $1x10^{-7}$ Ω-m. Resistance is equal to (resistivity x length) / area. Using the crack length number above and assuming a crack area equal to 75% of the crack area above (i.e. 25 % contact), the resulting resistance of the remaining attached solder in the crack is:

$$R = (1x10^{-7} \text{ x } 0.1x10^{-6}) / (.25 \text{ x } 7.068583x10^{-8}) = 5.66 \text{ x } 10^{-7} \text{ ohms}$$

The impedance, Z, is given by equation 3:

$$Z = 1/\text{Sqrt}\,((1/R)^2 + (1/X_C + 1/X_L)^2) \qquad \text{(Equation 3)}$$

After calculating capacitive reactance and resistance above, one can ignore the inductive reactance. So, plugging in the number for resistance and capacitive reactance (and using 75% of that value), one can estimate Z as 5.66 x 10^{-7} ohms as the very low resistance dominates.

$$\text{Conductance } G = 1/R = 1.767 \text{ x } 10^6 \qquad \text{(Equation 4)}$$

And

$$\text{Admittance } Y = 1/Z = 1.767 \text{ x}10^6 \qquad \text{(Equation 5)}$$

One can now estimate the phase angle, using equation 6, and relate this to the S-parameters in this testing. The phase angle will not change even if only a small portion of the joint is still attached; the phase angle will change if there is a complete break.

$$\text{Cosine } \varphi = G/Y = 1, \text{ so } \varphi = 1. \qquad \text{(Equation 6)}$$

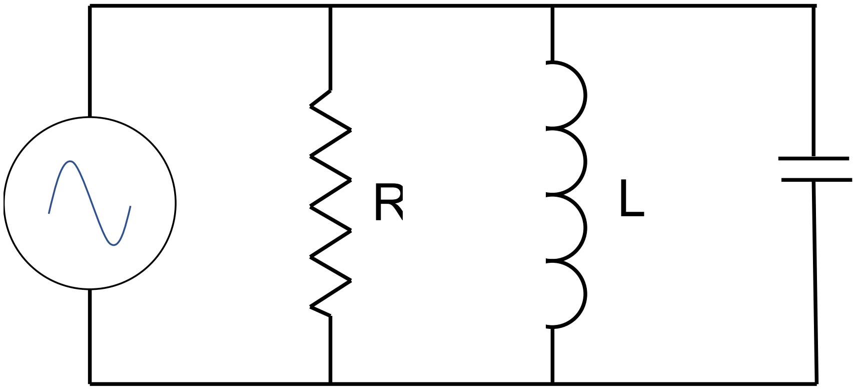


Figure 11 Solder Crack Equivalent Circuit

The insertion loss (S21) will only see degradation of the signal and the changes to the circuit as detailed above will be very small compared to the losses in the entire circuit. At very high frequencies, low resistance dominates the crack properties, and will not be seen until the crack opens completely so that the resistance is substantially increased. It generally takes multiple cycles after a DC open to create a significant separation of the solder joint. This, and the polling frequency, which is by necessity much lower than for DC monitoring, may help explain why the RF data suggests that the solder joints fail later.

The theoretical analysis also shows why failures would be identified at the same cycle if the analysis is done using insertion loss, return loss, or unwrapped phase angle as the measurand. Return loss will see a change in impedance. The analysis above shows impedance is overwhelmingly driven by the very low resistance of even a small remaining portion of the solder joint. After the crack opens and the resistance becomes very high, then the capacitive reactance drives the impedance and it will go up. Since the impedance of the crack is dominated by the low resistance and the conductance is also dominated by the low resistance, the phase angle change from the crack is effectively zero until the crack opens, when both impedance and conductance change. The theoretical analysis agrees with the results obtained in this work; RF monitoring will not see a failing solder joint until during, or even after, DC monitoring of the same solder joint, as shown by the statistical analysis (Weibull plot, Figure 7)

Conclusions

The following are the conclusions from this work:

1) The historical DC based solder joint failures and solder joint reliability evaluations methods are also acceptable for high frequency circuits.
2) Failures in high frequency circuits from solder joint cracking are expected to be detected simultaneously or even after failures are detected with DC methods, as the resulting capacitance of the solder joint crack is effectively a short circuit for high frequency circuits.
3) The capacitance and associated capacitive reactance, formed by a partial crack in a solder joint is so much larger than the very small resistance of a tiny remaining amount of intact solder joint. The ultimate effect on the circuit is unmeasurable until the crack is fully open. The low resistance dominates the circuit.

Recommendations for future work

The analysis reported in this paper was based on a 0.5mm pitch Waferscale CSP device with very small solder joints. While the results are expected to be reproducible, a more extensive study, especially with devices with larger solder joints, or much smaller solder joints, would be required to verify the universality of the observations and conclusions.

For the proposed validation studies, a dedicated VNA is recommended for each net in order to avoid complications in the data analysis from the use of multiplexers and high frequency switches which complicate and add to measurement variations and uncertainties.

References

1) Bin, Y., Yudong, L. and Ming, W. (2013), "Solder joint degradation and detection using RF impedance analysis", Soldering & Surface Mount Technology, Vol. 25 No. 1, pp. 25-30. https://doi.org/10.1108/09540911311294579

2) Kwon et al. "Early Detection of Interconnect Degradation by Continuous Monitoring of RF Impedance," in IEEE Transactions on Device and Materials Reliability, vol. 9, no. 2, pp. 296-304, June 2009

3). Azarian et al., "The role of impedance control in early detection of interconnect degradation using time domain reflectometry," *2012 IEEE 16th SPI*, Sorrento, 2012, pp. 21-24.

4) "Detection of solder joint degradation using RF impedance analysis", D Kwon, MH Azarian, MG Pecht - 2008 58th Electronic Components and Technology, 2008).)

5) M. H. Azarian, et al., "An analytical model of the RF impedance change due to solder joint cracking," 2011 IEEE 15th Workshop on Signal Propagation on Interconnects (SPI), Naples, 2011, pp. 89-92. doi: 10.1109/SPI.2011.5898847

6) C. Okoro, P. Kabos, J. Obrzut, K. Hummler and Y. S. Obeng, "Accelerated Stress Test Assessment of Through-Silicon Via Using RF Signals," in IEEE Transactions on Electron Devices, vol. 60, no. 6, pp. 2015-2021, June 2013, doi: 10.1109/TED.2013.2257791.

7) "Understanding the Pre-Failure Thermo-Mechanical Issues in Electromigration of TSV Enabled 3D ICs" C. E Sunday, et al., ECS Transactions, Volume 77 (2), 231st ECS Meeting, May 28, 2017–June 1, 2017, New Orleans, LA,

8) "Microwave Monitoring of Atmospheric Corrosion of Interconnects", Amoah P. K et al., ECS J. Solid State Sci. Technol. 2018 volume 7, issue 12, N143-N149, doi: 10.1149/2.0181812jss

9) "Microwave-Based Metrology Platform Development: Application of Broad-Band RF Metrology to Integrated Circuit Reliability Analyses", You L. et al, ECS Transactions, 61 (6) 113-121 (2014), doi: 10.1149/06106.0113ecst

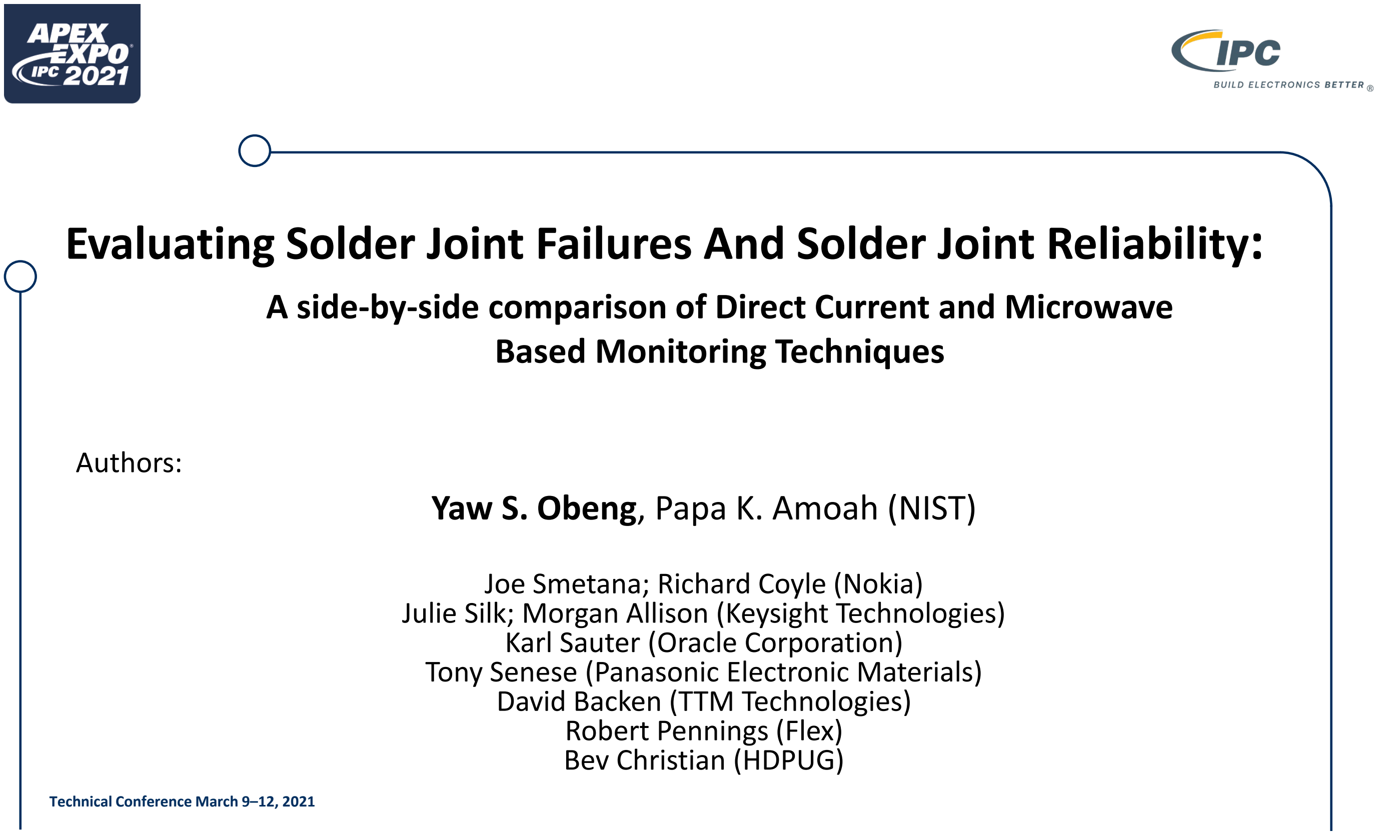
10) "Microwave-Based Metrology Platform Development: Application of Broad-Band RF Metrology to Integrated Circuit Reliability Analyses", L. You, et al., ECS Transactions: Moore-Than-More 2, 225th Electrochemical Society, May 4-9, 2014, Orlando, FL

11) "Signal Integrity Analysis of TSV Enabled 3-D IC", Yaw Obeng and David Love, HDPUG Project Report, archived on HDPUG TSV Signal Integrity website, 2016.

12) Pozar, David M. Microwave Engineering. United States, Wiley, ISBN:9781118213636, 1118213637, 2012.

13) Kaushik, L., Azarian, M.H. & Pecht, M. Effect of different lubricant films on contact resistance of stationary separable gold-plated electrical contacts. J Mater Sci: Mater Electron 30, 14368–14381 (2019). https://doi.org/10.1007/s10854-019-01806-y

This paper was first presented at the 2021 IPC Apex Expo Technical Conference and published in the 2020 Technical Conference Proceedings.



Evaluating Solder Joint Failures And Solder Joint Reliability:

A side-by-side comparison of Direct Current and Microwave Based Monitoring Techniques

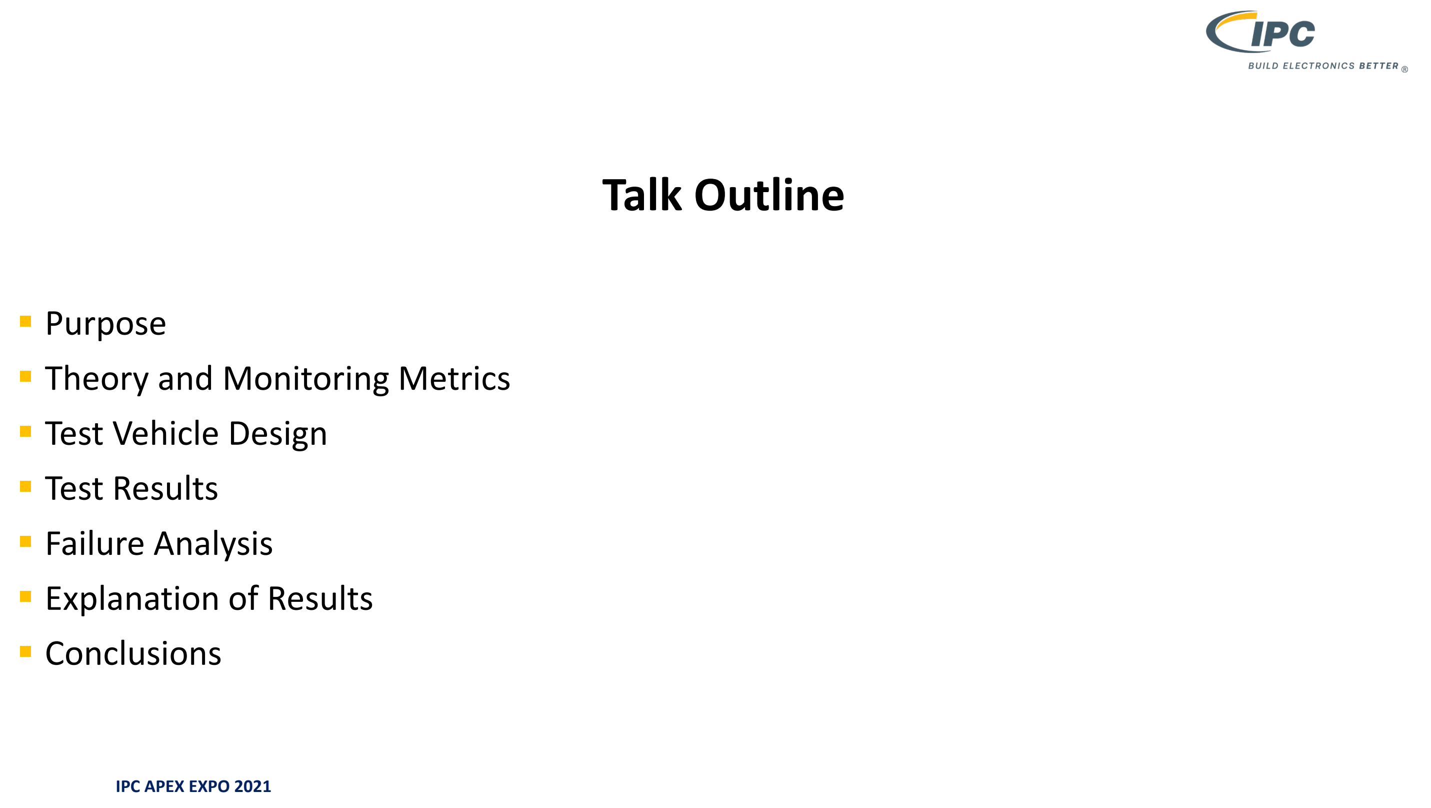
Authors:

Yaw S. Obeng, Papa K. Amoah (NIST)

Joe Smetana; Richard Coyle (Nokia)
Julie Silk; Morgan Allison (Keysight Technologies)
Karl Sauter (Oracle Corporation)
Tony Senese (Panasonic Electronic Materials)
David Backen (TTM Technologies)
Robert Pennings (Flex)
Bev Christian (HDPUG)

Technical Conference March 9–12, 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

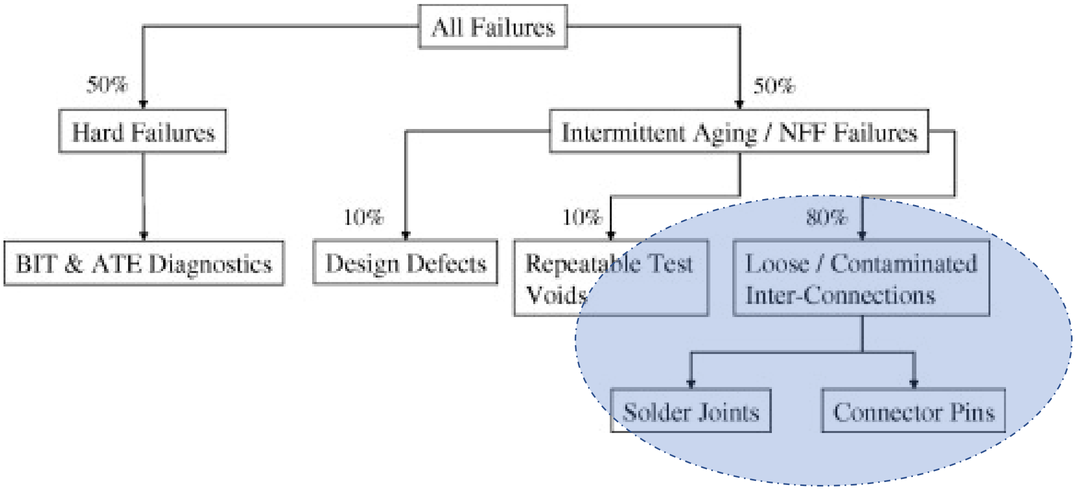


Talk Outline

- Purpose
- Theory and Monitoring Metrics
- Test Vehicle Design
- Test Results
- Failure Analysis
- Explanation of Results
- Conclusions

IPC APEX EXPO 2021

Hard and NFF Failures in Avionics

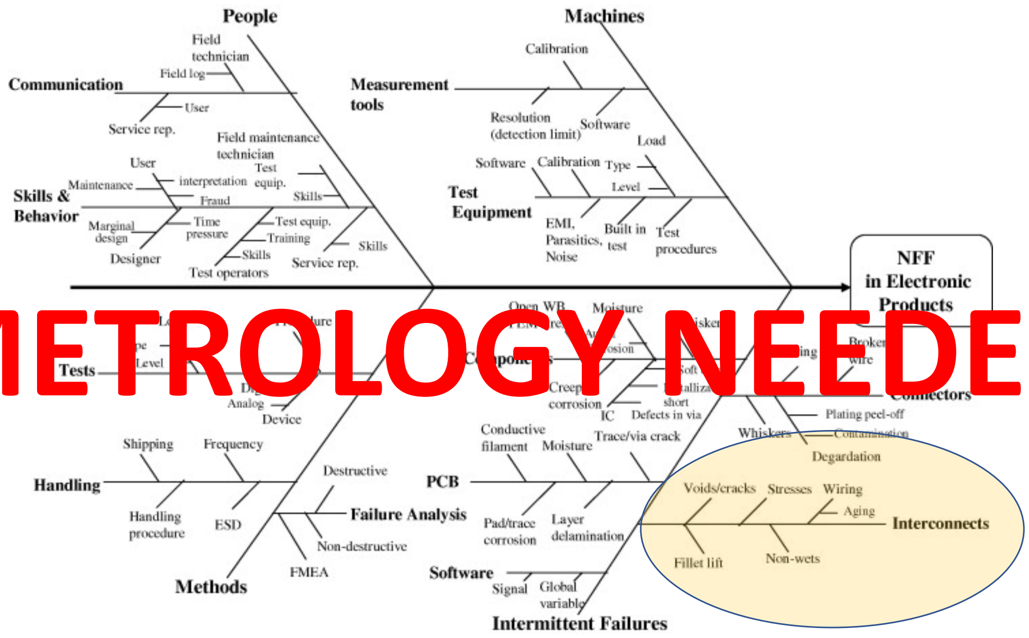


NFF failures costs the Defense Department (DoD) more than $2 billion annually

H. Q et al., Microelectronics Reliability, 2008, 48, 5, 663-674, doi: /0.1016/j.microrel.2008.02.003
C. Adams, Aviation Maintenance , 2014: Pages 26 -32.

IPC APEX EXPO 2021

Fish-Bone Diagram of NFF Issues



METROLOGY NEEDED

IPC APEX EXPO 2021

H. Q et al., Microelectronics Reliability, 2008, 48, 5, 663-674, doi: /0.1016/j.microrel.2008.02.003

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Purpose of the Work

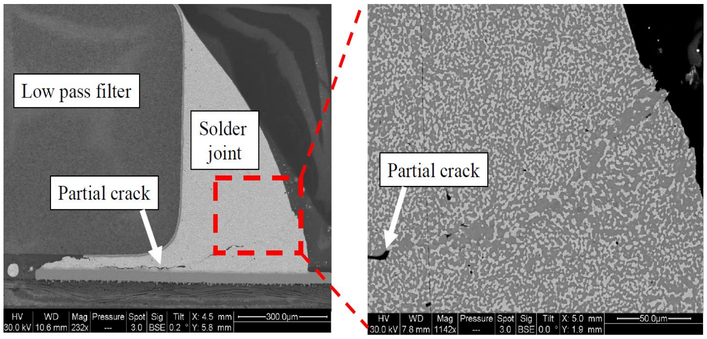
Determine whether RF-Based Is More Sensitive Than DC-Resistance Metrology in Detecting Solder Joints Degradation.

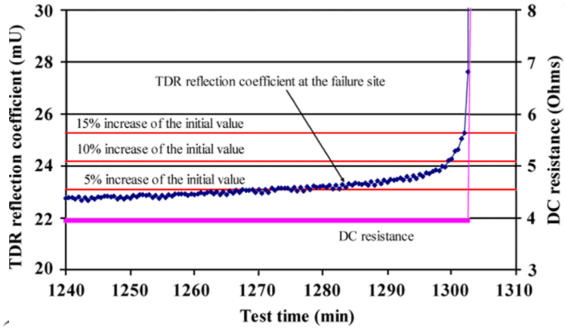
The goals:

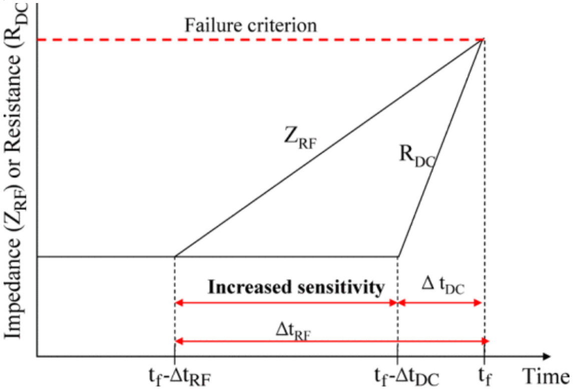
- Understand the impact of defects on RF- and DC signal losses in I/O assemblies
- Use the signal loss phenomena to detect incipient defect formation, and determine the onset of performance limitation
- Statistically determine the differences between DC- and RF- detection capability for solder joints in thermal cycling
- Perhaps change our definition of the “failed” solder joint
- Potentially explain many No Failure Found (NFF) field returns

Obeng, Yaw S.; Amoah, Papa. ”Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques.” Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

RF Signal Losses Due to Pre-catastrophic Solder Joint Failure

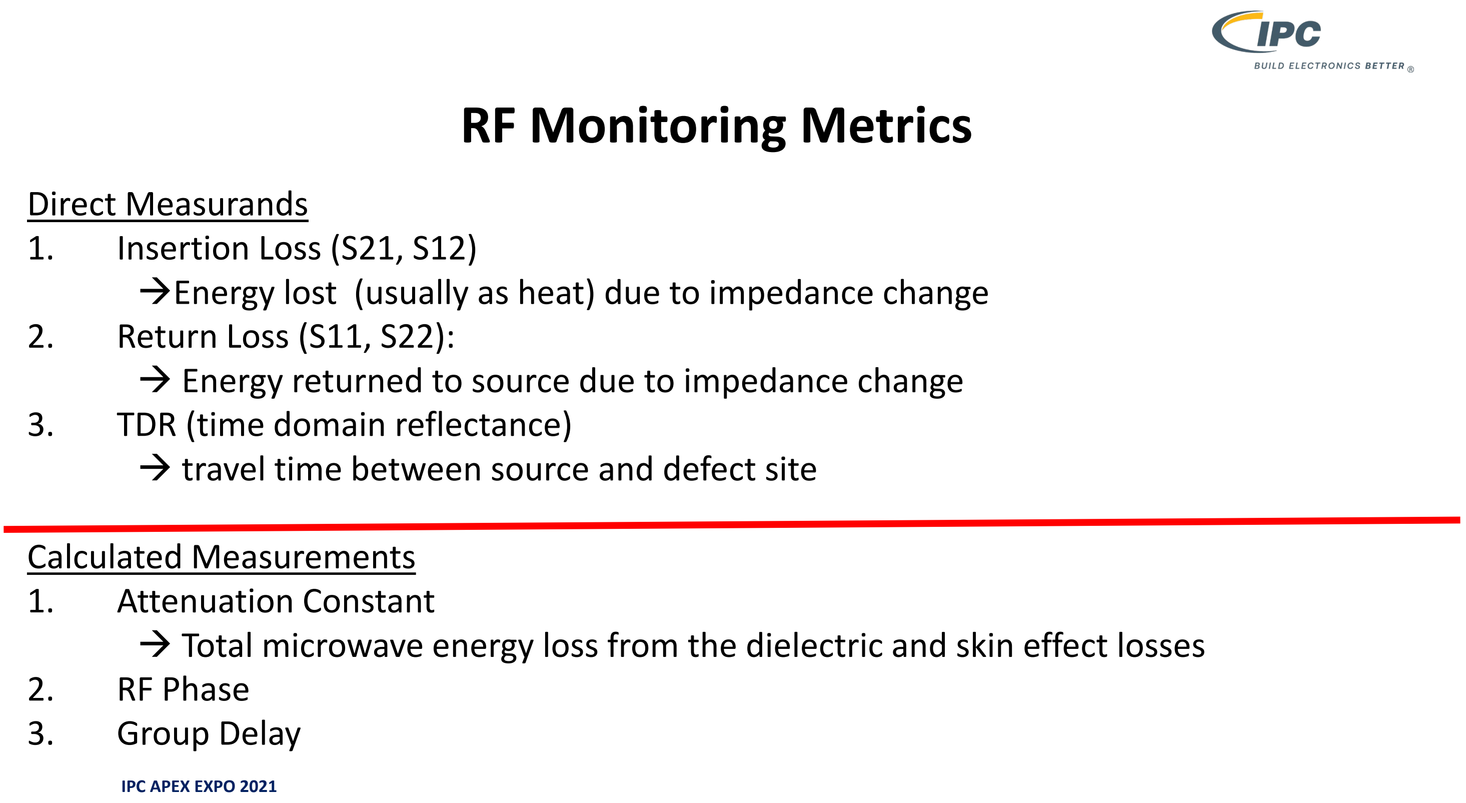

Low pass filter
Solder joint
Partial crack
Partial crack
HV 30.0 kV WD 10.6 mm Mag 232x Pressure --- Spot 3.0 Sig BSE Tilt 0.2 ° X: 4.5 mm Y: 5.8 mm 300.0µm
HV 30.0 kV WD 7.8 mm Mag 1142x Pressure --- Spot 3.0 Sig BSE Tilt 0.0 ° X: 5.0 mm Y: 1.9 mm 50.0µm





D. Kwon, et al., IEEE Transactions on Device and Materials Reliability, vol. 9, no. 2, pp. 296-304, June 2009

IPC APEX EXPO 2021



RF Monitoring Metrics

Direct Measurands

1. Insertion Loss (S21, S12)
   → Energy lost (usually as heat) due to impedance change
2. Return Loss (S11, S22):
   → Energy returned to source due to impedance change
3. TDR (time domain reflectance)
   → travel time between source and defect site

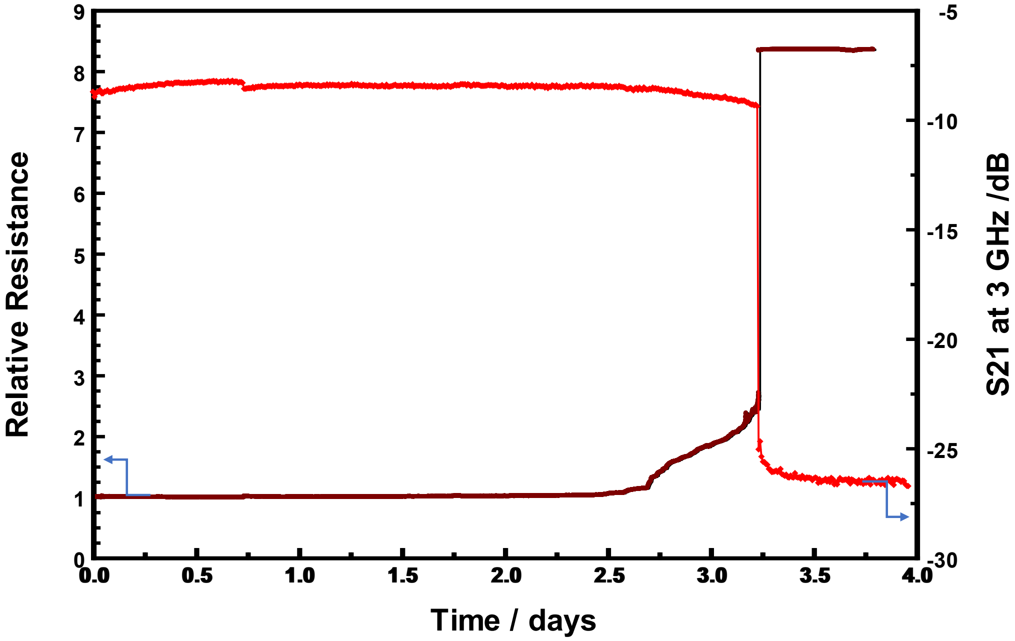
Calculated Measurements

1. Attenuation Constant
   → Total microwave energy loss from the dielectric and skin effect losses
2. RF Phase
3. Group Delay

IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

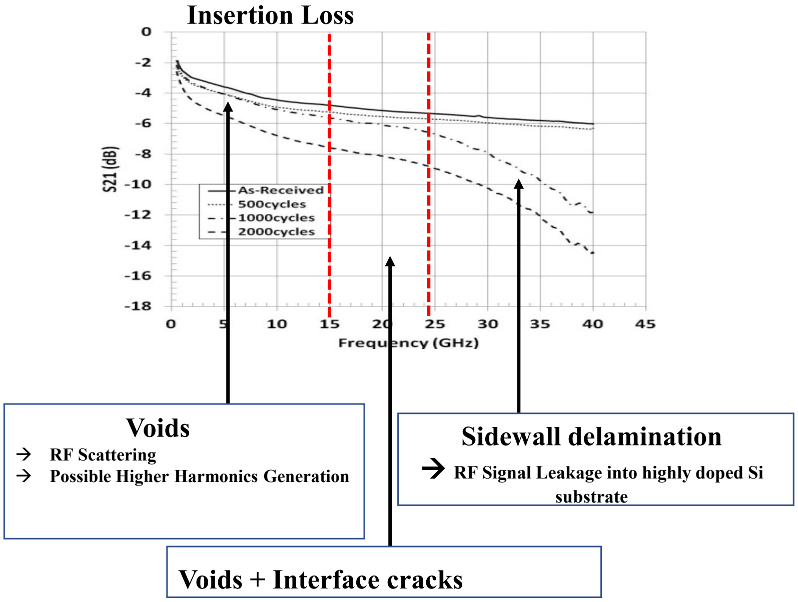
CORRELATION BETWEEN RF INSERTION LOSS (S21) AND DC REISTANCE (R_{DC})



IPC APEX EXPO 2021

Christopher E. Sunday; et al., Journal of Applied Physics 2017, 122, DOI: 10.1063/1.4992135
Papa K. Amoah, et al., Microwave Monitoring of Atmospheric Corrosion of Interconnects. ECS Journal of Solid-State Science and Technology 2018, 7:12, N143-N149.

Insertion Loss (S21) as Thermomechanical Reliability Probe



Voids

- → RF Scattering
- → Possible Higher Harmonics Generation

Sidewall delamination

→ RF Signal Leakage into highly doped Si substrate

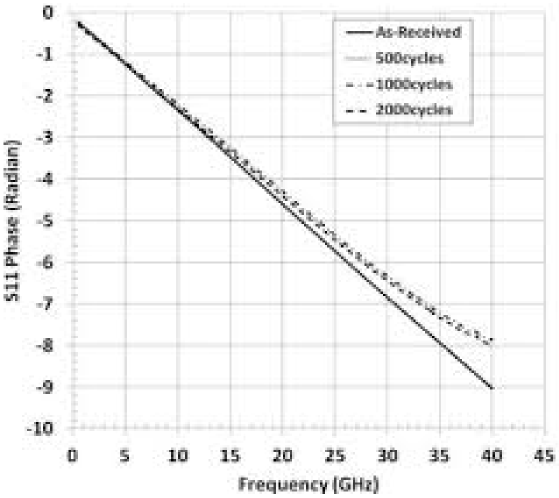
Voids + Interface cracks

IPC APEX EXPO 2021

Okoro et al. ,"Accelerated Stress Test Assessment of Through-Silicon Via Using RF Signals", IEEE TRANSACTIONS ON ELECTRON DEVICES, 60 (6), J, 2013 pp 2015-2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Return Loss (S11) Phase as Thermomechanical Reliability Probe



Unwrapped Phase Angle

$$\mathit{Phase\ Shift}\ \phi = \arctan \frac{1}{2\pi fRC}$$

IPC APEX EXPO 2021

Okoro et al. ,"Accelerated Stress Test Assessment of Through-Silicon Via Using RF Signals", IEEE TRANSACTIONS ON ELECTRON DEVICES, 60 (6), J, 2013 pp 2015-2021

S-parameters Afford Complete Suite Of Electrical Parameters

$$R=\mathrm{Real}\left(Z_0\frac{(1+S_{11})(1+S_{22})-S_{12}S_{21}}{2S_{21}}\right)$$

$$L=\frac{\mathrm{Imag}}{\omega}\left(Z_0\frac{(1+S_{11})(1+S_{22})-S_{12}S_{21}}{2S_{21}}\right)$$

$$G=\mathrm{Real}\left(\frac{(1+S_{11})(1-S_{22})+S_{12}S_{21}-2S_{21}}{Z_0((1+S_{11})(1+S_{22})-S_{12}S_{21})}\right)$$

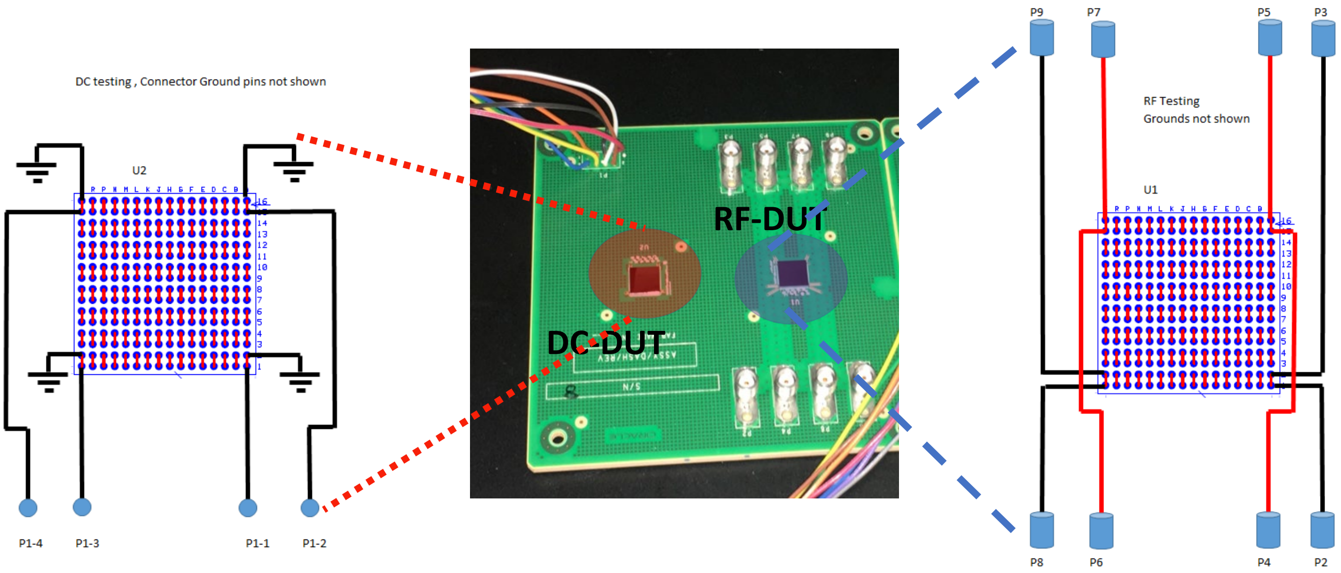
$$C=\frac{\mathrm{Imag}}{\omega}\left(\frac{(1+S_{11})(1-S_{22})+S_{12}S_{21}-2S_{21}}{Z_0((1+S_{11})(1+S_{22})-S_{12}S_{21})}\right)$$

IPC APEX EXPO 2021

Device Wiring for In Situ Monitoring

DC-DUT

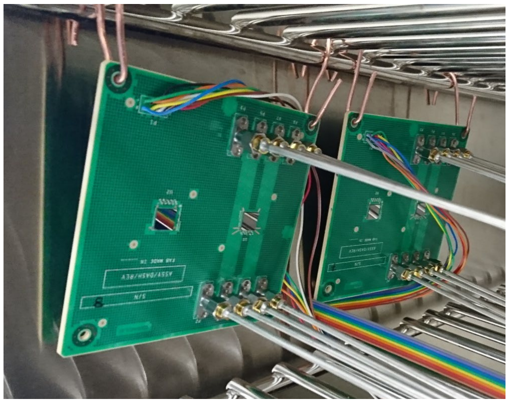
RF-DUT

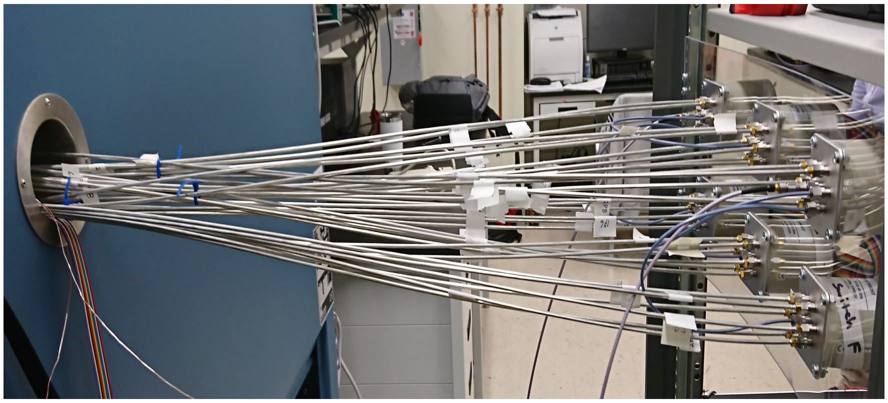


IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Boards Mounting and Connections to Switches

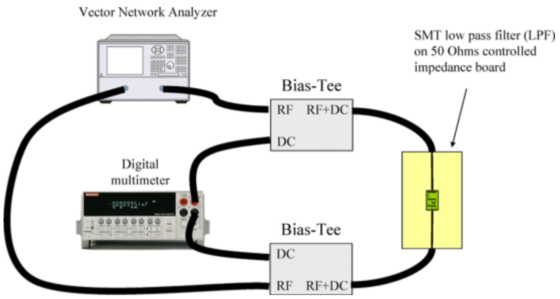




IPC APEX EXPO 2021

In-Situ Monitoring: 2 Ways to Run the Test

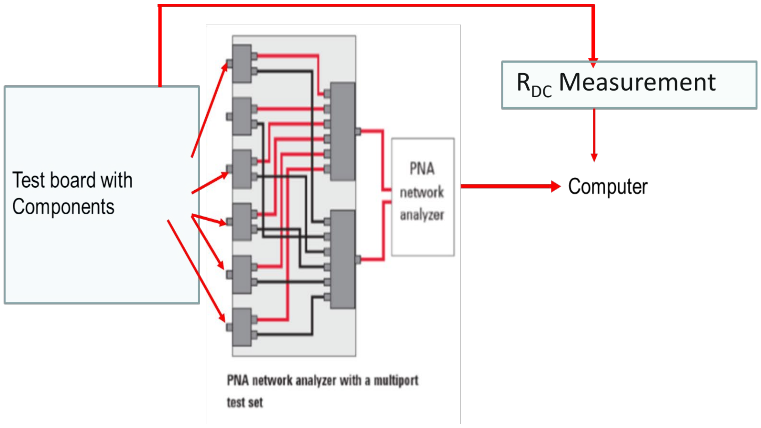
Single Device Tested with 2 Circuits (RF and DC)



D. Kwon, M. H. Azarian and M. Pecht, IEEE Transactions on Device and Materials Reliability, vol. 9, no. 2, pp. 296-304, June 2009

IPC APEX EXPO 2021

Multiple Devices Tested with 2 Circuits (RF and DC)

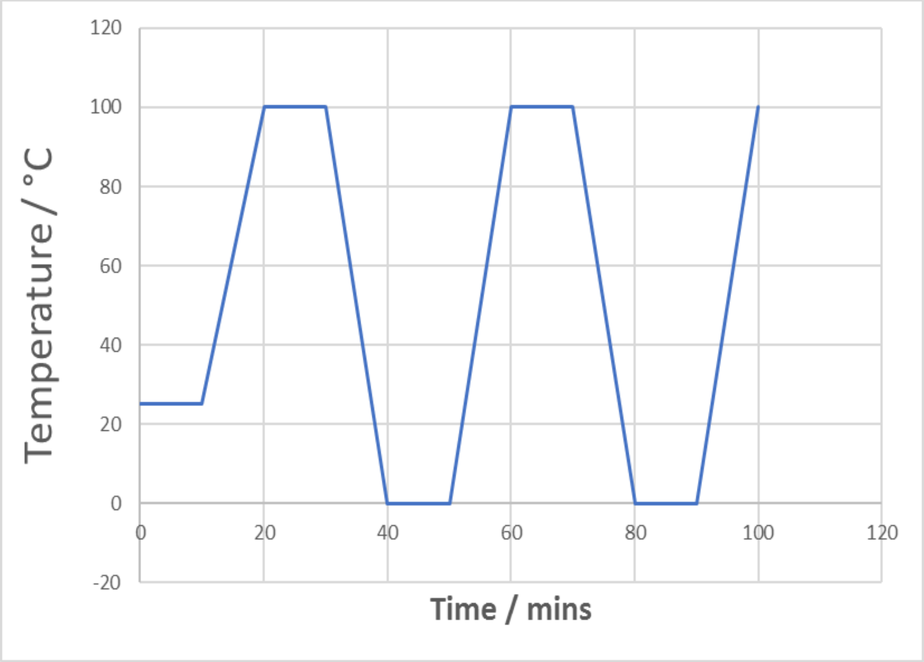

Test board with Components

R_{DC} Measurement

Computer

Thermal Cycle Profile

IPC BUILD ELECTRONICS BETTER ®

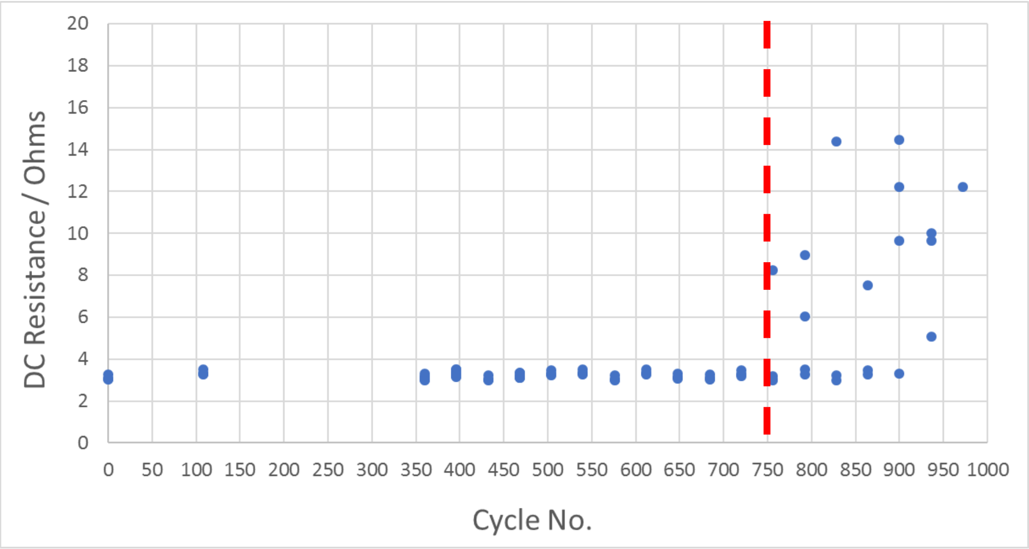


1000 cycles of :
10 min ramp-up
10 min soak
10 min ramp-down

IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

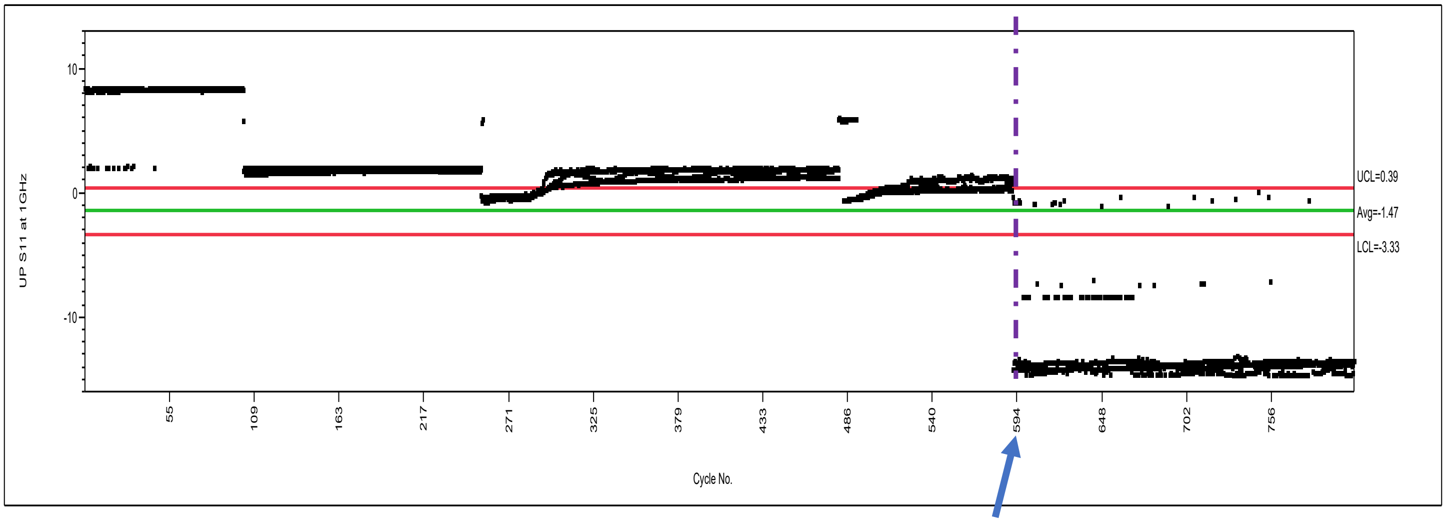
Typical DC-Resistance Evolution



IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Typical RF Phase Angle Evolution



S11 Phase Angle at 1 GHz

RF Failure Onset

IPC APEX EXPO 2021

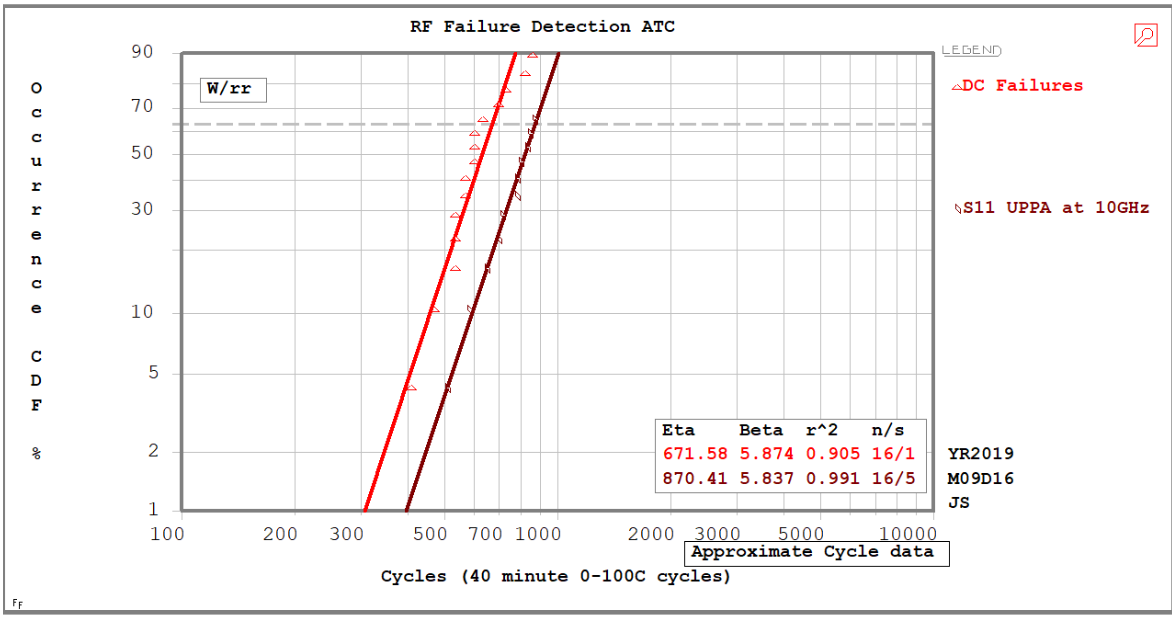
Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Potential Sources of Uncertainties in the RF Data

- Uncompensated temperature dependent gradients along semi-rigid cable characteristics during the thermal cycling.
  - This could add quite a bit of noise in the measurement.
  - We were unable to correlate noise in RF data with temperature transitions during the thermal cycling
- Mechanical aging of metallic contacts in the RF switch / relay network
  - Possible fretting of contact surfaces?

IPC APEX EXPO 2021

R_{DC} vs RF Failure Rates

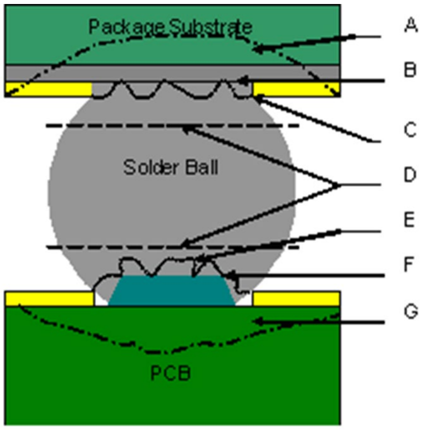


Eta	Beta	r^2	n/s
671.58	5.874	0.905	16/1
870.41	5.837	0.991	16/5

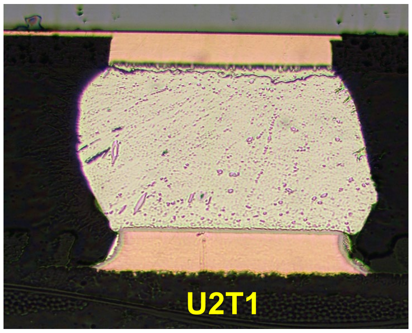
IPC APEX EXPO 2021

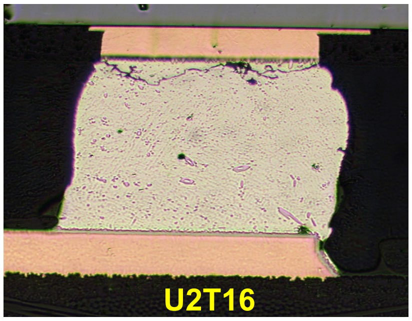
Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Physical Failure Analysis



A Package Pad Lift/Crater
B Pkg Base Metal/IMC Interface Fracture
C Pkg IMC/Solder Interface Fracture
D Bulk Solder Fracture
E PCB IMC/Solder Interface Fracture
F PCB Solder pad/IMC Interface Fracture
G PCB Pad Lift/Crater





IPC APEX EXPO 2021

"METHODOLOGY TO CHARACTERIZE PAD CRATERING UNDER BGA PADS IN PRINTED CIRCUIT BOARDS" in the Proceedings of the Pan Pacific Microelectronics Symposium, Kauai, Hawaii, January 22 – 24, 2008.

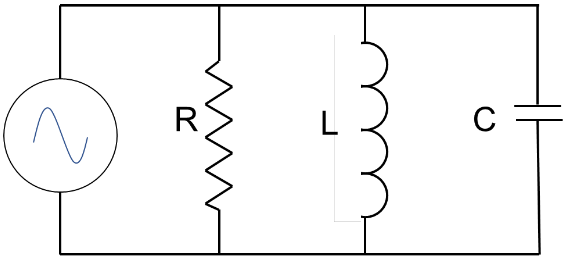
Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Summary of Observations

- The DC failures preceded the RF failures
  - R_{DC} monitoring technique predicted shorter lifetimes than the RF monitoring.
- The return loss (S11) metrics (i.e., Amplitude and Unwrapped phase (UP) angle) afforded the identical failure onset time at both frequencies.
- Solder joint failed by fracture at / near the IMC-Solder interface.

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

Why the Difference Between R_{DC} and RF Lifetime Results?



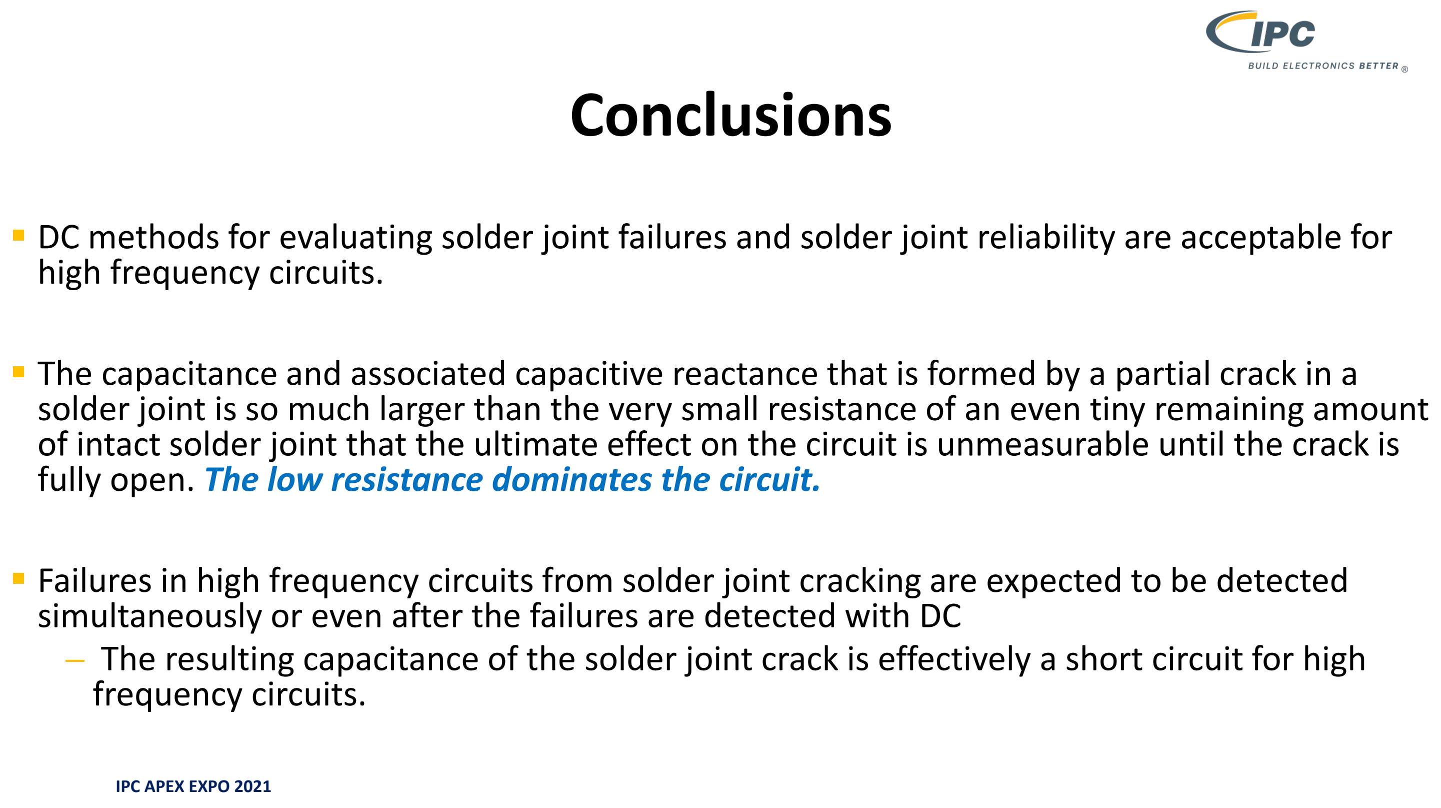
Solder Crack Equivalent Circuit

Key Point *: As long as there is still a Connection in the crack, R dominates the impedance*
(akin to a Passive High Pass Filter)

- X_L so small it can be ignored
- X_c = 1/(2πfC) = for a crack ~ 2.5ohms at 10GHz
- R crack = Extremely small (~5.7 x 10^{-7} ohms) until solder joint opens -
- Z = 1/Sqrt $((1/R)^2 + (1/XC + 1/XL)^2)$ = 5.66x10^{-7} ohms – R dominates

IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

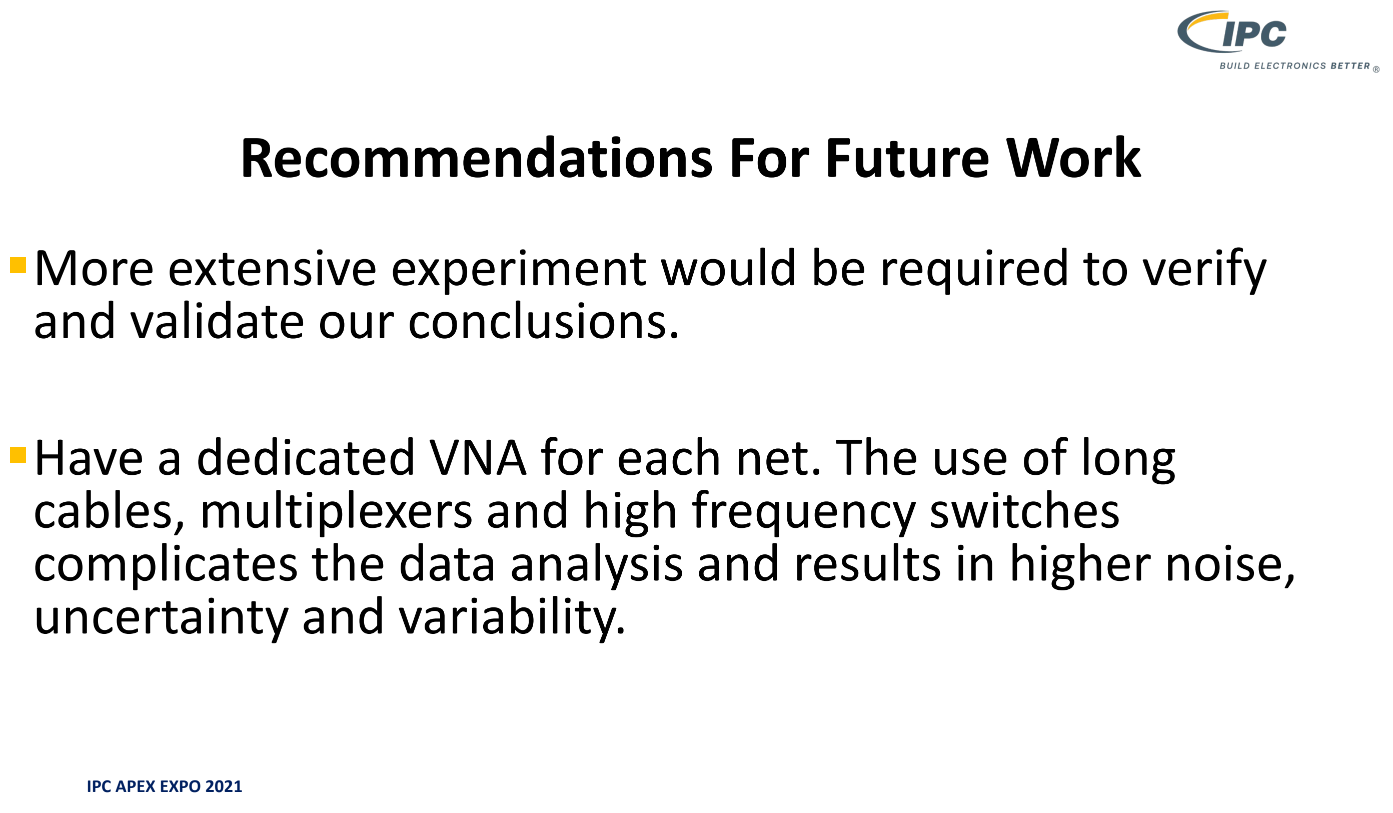


Conclusions

- DC methods for evaluating solder joint failures and solder joint reliability are acceptable for high frequency circuits.
- The capacitance and associated capacitive reactance that is formed by a partial crack in a solder joint is so much larger than the very small resistance of an even tiny remaining amount of intact solder joint that the ultimate effect on the circuit is unmeasurable until the crack is fully open. ***The low resistance dominates the circuit.***
- Failures in high frequency circuits from solder joint cracking are expected to be detected simultaneously or even after the failures are detected with DC
  - The resulting capacitance of the solder joint crack is effectively a short circuit for high frequency circuits.

IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.



Recommendations For Future Work

- More extensive experiment would be required to verify and validate our conclusions.
- Have a dedicated VNA for each net. The use of long cables, multiplexers and high frequency switches complicates the data analysis and results in higher noise, uncertainty and variability.

IPC APEX EXPO 2021

Obeng, Yaw S.; Amoah, Papa. "Evaluating Solder Joint Failures and Solder Joint Reliability: A Side-by-Side Comparison of Direct Current and Microwave Based Monitoring Techniques." Presented at IPC Apex Expo 2021. March 08, 2021 - March 12, 2021.

NIST SP 1283
August 2022

Proceedings 2021 ASME International Manufacturing Science and Engineering Conference
MSEC2021
June 21-25, University of Cincinnati, USA

MSEC2021-63892

AUTO-CALIBRATION FOR VISION-BASED 6-D SENSING SYSTEM TO SUPPORT MONITORING AND HEALTH MANAGEMENT FOR INDUSTRIAL ROBOTS

Guixiu Qiao
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

Guangkun Li
Johns Hopkins University Applied Physics
Laboratory, Maryland, USA

KEY WORDS

Advanced Sensing, Vision-based System Calibration, Prognostics and Health Management (PHM), Industrial Robot, Accuracy Degradation

ABSTRACT

Industrial robots play important roles in manufacturing automation for smart manufacturing. Some high-precision applications, for example, robot drilling, robot machining, robot high-precision assembly, and robot inspection, require higher robot accuracy compared with traditional part handling operations. The monitoring and assessment of robot accuracy degradation become critical for these applications. A novel vision-based sensing system for 6-D measurement (six-dimensional x, y, z, yaw, pitch, and roll) is developed at the National Institute of Standards and Technology (NIST) to measure the dynamic high accuracy movement of a robot arm. The measured 6-D information is used for robot accuracy degradation assessment and improvement. This paper presents an automatic calibration method for a vision-based 6-D sensing system. The stereo calibration is separated from the distortion calibration to speed up the on-site adjustment. Optimization algorithms are developed to achieve high calibration accuracy. The vision-based 6-D sensing system is used on a Universal Robots (UR5) to demonstrate the feasibility of using the system to assess the robot's accuracy degradation.

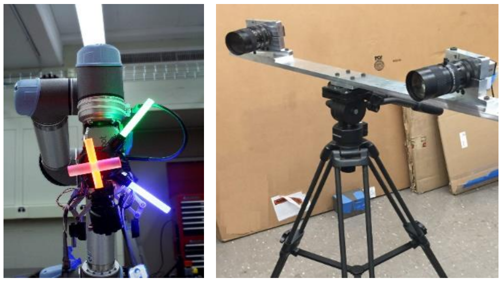
INTRODUCTION

Industrial robot systems are very complex, containing sub-systems and components that interact within manufacturing work cells. It is a challenge to determine their specific influences on performance when an unexpected break down happens in a work cell [1, 2]. There is increasing interest to enable the ability to leverage prognostic data to monitor the system's health status [3, 4]. Data analysis may generate actionable intelligence in maintenance plan optimization. The monitoring of robot health conditions may also help to detect potential faults and failures. It helps to eliminate the unexpected maintenance shutdowns that are expensive.

To monitor the health condition of an industrial robot, low-level data can be extracted from robot controllers. Information from controllers may include joint current, positions, velocities, accelerations, etc. The controller's low-level data can be used to capture and monitor the abnormal changes of signals [5]. But this method is suitable for repeating operations, where signals usually have fixed patterns at normal conditions. Moreover, some non-geometric errors, for example, squareness errors between axes and deflections of the structure cannot be reflected from joint data. An individual joint's data cannot determine the overall robot's health condition, that is, the accuracy degradation of the robot's tool center position (TCP). To monitor a robot's absolute accuracy changes, a sensor that can measure the 6-D information of the robot TCP is needed.

Many sensing systems in the market can measure 6-D information, including laser trackers [6], theodolite measurement devices [7], probing coordinate measuring machines (CMM) [8, 9], and optical tracking systems [10, 11]. Laser trackers and theodolites have the line-of-sight issue for moving objects. CMM measurement is slow, not suitable to measure the robot arm's dynamic movements. Also, laser trackers and CMM measurements are expensive. Optical tracking systems are vision-based instruments. The major challenge to optical tracking lies in measurement accuracy. Since they use infrared (IR) cameras and reflective spheres as targets, ambient light has a strong influence on measurement uncertainties [10]. A novel vision-based 6-D sensing system (patent pending) has been developed at the National Institute of Standards and Technology (NIST). The system contains a smart target (Fig. 1a)) and a vision-based measurement instrument (Fig 1b)). The smart target has three-color light pipes. It is motor-driven and can constantly rotate toward the measurement instrument. There is an orientation sensor mounted on the elevation shaft. An initial pose is defined by the user. When the smart target rotates away from the original pose, motors on the azimuth and elevation shaft rotate the elevation shaft back to its

Qiao, Guixiu; Li, Guangkun. "Auto-Calibration for Vision-Based 6-D Sensing System to Support Monitoring And Health Management for Industrial Robots." Presented at 2021 International Manufacturing Science and Engineering Conference (MSEC 2021). June 21, 2021 - June 25, 2021.



a) Smart target (mounted on the end effector of UR5)

b) Vision-based measurement instrument

Figure 1. Vision-based 6-D measurement

original pose driven by the orientation sensor's feedback. The novel design enables high accuracy (within 0.1mm) and high speed (at a minimum of 30 Hz) in measurement. The design of the smart target system is detailed in [12, 13]. This paper focuses mainly on the automatic calibration technology to achieve the high accuracy of the vision-based 6-D sensing system.

This paper is organized as follows: the next section describes the automatic distortion calibration development. Then refinement algorithms are discussed that use iteration calculations to handle tilt images and imperfect calibration boards. Next, the verification method for the calibration result is described. The last section presents the establishment of an automatic stereo calibration procedure.

CAMERA AUTOMATIC CALIBRATION DEVELOPMENT

Traditional camera calibration methods compute camera intrinsic parameters from a set of target features with known geometries. The most common target is a checkerboard. Calibration usually requires users to select corners and areas on the checkerboard images manually. The manual process has several problems:

1) Time consuming. Users must manually click four corners on each checkerboard image to define the region of interest. When there are multiple images, this process is very time-consuming.

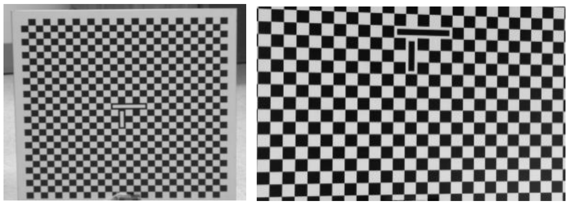
2) Prone to error. The corner detection relies on the accuracy of the user clicking that is prone to error.

3) Difficulty for stereo camera calibration. Using the dual camera as an example, stereo camera calibration needs to find correspondence between the left and right camera. Users either have to make sure both cameras contain the full checkerboard image, or count and align the same corners for both left and right images manually. This process is tedious and time-consuming.

A method for automatic camera calibration was developed. This method doesn't involve user clickings to define the region of interest. Images are processed in batch. The key to the automatic method is to enable auto-counting and auto-alignment. As shown in Fig. 2a), the calibration board used at NIST has two long rectangle makers in the center. The special markers are utilized as the indicator of the checkboard's orientation and the center position. If the marker features are correctly detected, any corner in the image can find the correspondence to the center. Users may customize their markers on their checkerboard, for example, adding extra dots or special shapes as indicators. The automatic procedure for calibration board corner detection is as follows.

1. Obtaining reverse images

The original image of the calibration board has white rectangles as the markers shown in Fig 2a). If the white rectangle is used as the indicator, false detections have a large chance to happen because the edge of the calibration board is white. To solve this problem, a reverse calculation is implemented to reverse the white rectangle to black color, as shown in Fig. 2b) Then the contours of the two black rectangles are detected.



a) Original calibration board b) Reverse image of calibration board

Figure 2. Calibration board

2. Obtaining binarized image

Images need to be filtered before further processing based on a threshold to obtain a binarized image. Two methods are developed to identify a proper threshold for the image. The first method is the adaptive thresholding based upon the Otsu Binarization method [14]. This is a well-established threshold method and we used an algorithm from the OpenCV library. This method is less robust and prone to false thresholding but runs faster. An improved method is developed by adding the calculation of image histogram, histogram smoothing, calculation of histogram gradient, and smart gradient divide thresholding. This new method is more robust compared with the first method. Because the algorithm is more complex, the calculation takes more time.

3. Squares and rectangles detection

Given the binary images, an algorithm is needed to detect the contours of black blobs. The method to find the contours can be simplified to finding four lines by using iteration and area approximation. The result is shown in Fig. 3. The detected long

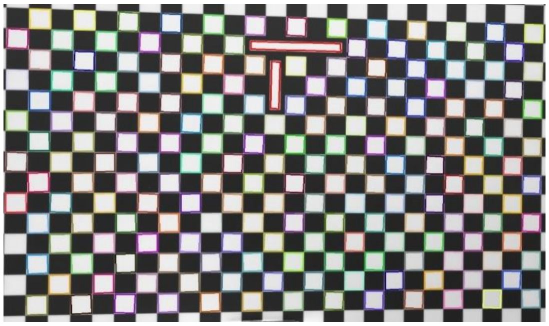


Figure 3. Illustration of the valid contours

rectangles are drawn in red color. Other detected squares are drawn in random colors. To remove noise and false detection, all the contours are arranged in a multi-level hierarchy-tree structure (parents and children) for further cleanup. In the end, the squares on the calibration board and the two long rectangles (drawn in red color in Fig. 3) are identified. The two long rectangles are further analyzed to determine the orientation. Corners are used as the 'anchor' to map the rest of the squares to a grid.

4. Recursive algorithm development for corner mapping

The left top corner of the long rectangle is treated as the anchor. The other square corners' positions are defined relative to the anchor. Then recursive searches are performed to find the nearest squares one-by-one. Grid coordinates are assigned to these square corners. By doing that, the so-called 'position awareness' is achieved. As a result, the corner correspondence from different images can be found automatically. Fig. 4 shows the assigned grid coordinates.

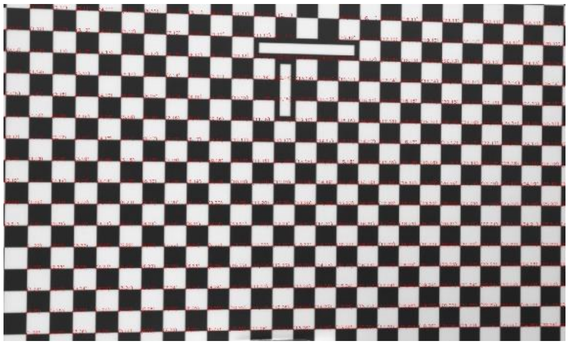


Figure 4. Labeled corners

5. Subpixel edge detection

The detected corners are passed through a subpixel edge detection algorithm. The algorithm runs in iterations to find the location of corners or radial saddle points in sub-pixel accuracy as shown in Fig. 5.

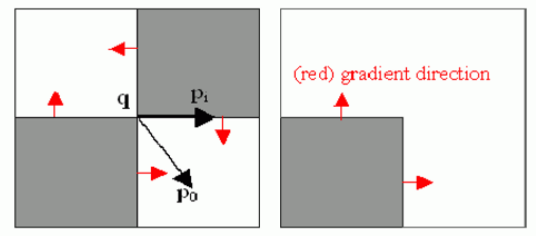
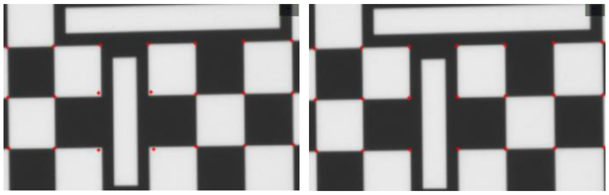


Figure 5. Subpixel edge detection algorithm from OpenCV library

During the development, multiple algorithms are implemented. It is found that the subpixel corner detection algorithms sometimes give the wrong location, as shown in Fig. 6a). The detected corners are marked with red crosses. When the corner is not a 'saddle' point, where near the anchor, the checkerboard conner becomes a white corner surrounded by the black background. The detected corner has deviated from the real corner locations. This problem is addressed by combining the current saddle point detection algorithm with a Harris corner localization approach [15]. Fig. 6b) shows that the results are more accurate using the new algorithm.



a) Corner detection with errors

b) Improved corner detection

Figure 6. Corner detection improvement

6. Alignment of the detected corners to calibration board coordinate frame

After the corners are detected precisely, an iterative recursive neighbor search is performed to align the corners to a grid. We define the left top corner of the calibration board to the coordinates (0,0). The right bottom corner coordinate is (26, 26) because the current calibration board has 26 x 26 grid size.

There are some requirements for taking calibration pictures. 1) Both of the rectangle markers must be seen in the image; 2) the calibration board cannot be rotated more than 45 degrees otherwise the algorithm may be confused by the location of the left top corner; 3) the calibration board (at least the two rectangle boards) occupies both cameras' field of view (FOV). Once the above procedures are followed, the calibration can be performed automatically and robustly. There is no need for a user to mouse-click hundreds of times which significantly reduces the problems caused by human errors.

CALIBRATION REFINEMENT APPROACH

Camera calibration accuracy is affected by the corner detection accuracy. Many calibration images are not fronto parallel (it means the calibration board is parallel to the image sensor). The titled calibration board image may cause non-linear distortion when localizing the subpixel corners [16]. A calibration refinement approach is developed to rectify the calibration images and refine the calibration with an iterative approach.

The refinement approach is described as follows.

- Corner detection: Detect the calibration board corners in the input images.
- First calibration: Use the detected corners to estimate camera parameters.
- Do iterations until convergence.
- Perform undistortion: Use the camera parameters to undistort and correct the input images to a frontal canonical pattern.
- Redo corner detection: Relocalize calibration pattern control points in the fronto parallel pattern.
- Re-project: Reproject the control points using the estimated camera parameters.
- Re-calibrate: Use the projected control points to re-fine the camera parameters.

Fig. 7 shows the original calibration images (left) and the fronto corrected images (right). Then the corner detection algorithms are applied to the fronto corrected images. The calibration procedure is performed for two iterations in our test. The calibration results show an improvement of about 5%. When the tilt angles of the calibration images are large, the improvement will be more significant with the benefit of the detection on fronto images.

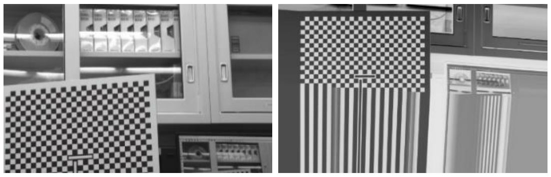


Figure 7. Original images and the fronto corrected images

Another algorithm is also implemented to handle the imperfect calibration board issue using an iterative approach [17]. Since the calibration board used at NIST is a high precision certified target, the existing test did not show significant improvement in calibration results. However, if users are using an imperfect calibration board, the developed algorithms are robust to handle the imperfect target condition.

VERIFICATION OF CALIBRATION RESULTS

After the calibration method is established, the calibration results need to be verified. A test method is developed using the straightness of line features to verify the distortion correction effects. An aluminate plate was machined to achieve surface flatness under 30 um. Three straight lines were precisely printed and applied on the top, middle, and bottom of the plane, as shown in the left picture of Fig. 8. The purpose of this test method is to evaluate the straightness of the lines that are detected by the camera after distortion correction. An unsuccessful calibration will result in curved lines with patterns. On the contrary, a successful calibration will result in a straight line. Thus, this is a good measure to verify the distortion correction of a calibration.

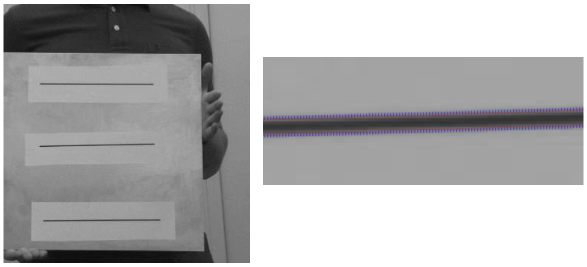


Figure 8. Image with lines used for distortion verification

Multiple line detection algorithms were tested to construct the line from the image. The traditional Hough line detection method has a problem in the test [18]. The way of finding the rising peak from the line edges does not perform very well in this case due to 1) saturation in the center of line; and 2) broken segment due to distortion. A robust line detection algorithm with sub-pixel level precision is desired. After tests, an algorithm based upon the approach by Trujillo-Pino et. al. [19] is implemented. The edge is located accurately based on orientation, intensity difference at both sides, subpixel position, and curvature. The detection results are shown in the right picture of Fig. 8. Once the edges are detected, we use the average location for both edges as the center and find the line.

Once lines are detected, a linear fit is performed to evaluate the straightness. Fig. 9 shows the linear fit of one of the lines before applying lens distortion correction to the image. Fig. 10 shows the residual errors of the line fitting before and after applying distortion correction. The horizontal axis draws the point numbers on the line. The tested line has about 600 points. The vertical axis is the residual error of each point in pixel units. Before the correction, as shown in Fig. 10a), there is a strong bending pattern and the residual error range is from -0.25 to 0.15 (in pixels). Fig. 10b) shows the improved results after correction. The residual error is significantly reduced. The residual error range is from -0.05 to 0.05 (in pixels) after distortion correction. The bending pattern is removed as well. The test demonstrates the effectiveness of the calibration results. This can also be used as a quick check of the vision-based system's accuracy before performing measurements.

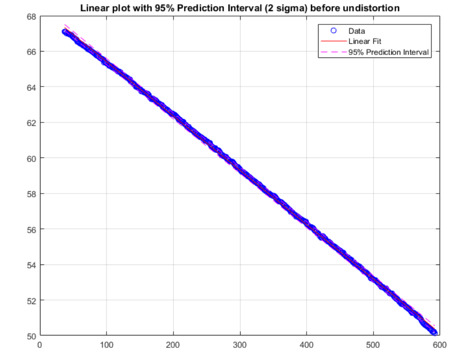
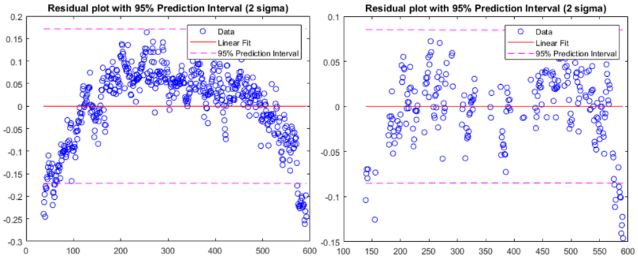


Figure 9. Linear fit of the line to evaluate



a) Before b) After

Figure 10. Residual error of the line fitting before and after distortion correction

ESTABLISH STEREO CALIBRATION PROCEDURE

In previous sections, an automatic camera calibration procedure for automatic corner detection and single camera calibration is presented. In this section, the stereo camera calibration procedure is presented.

The task of developing an automatic stereo camera calibration method is more challenging than the method for automatic single camera calibration. During stereo calibration, operators have to fit the calibration board into both cameras' field of view (FOV). Sometimes the calibration board looks smaller in the image, more tilted, and having more background, as shown in Fig. 11. These conditions present more problems. Stereo calibration requires more robust corner detection algorithms to handle the complicated background. Fig. 12 shows all the contours detected from the image. The two rectangles inside the red circles are what we used for the 'anchor'. Fig. 13 shows the contour detected for an image in the single camera calibration case, which shows the levels of challenges stereo auto-calibration is facing compared with Fig. 12. It is a challenging task to pick out all the calibration board corners from all the detected contours.

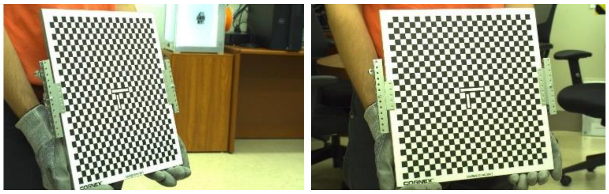


Figure 11. One pair of images from left and right cameras used for stereo calibration

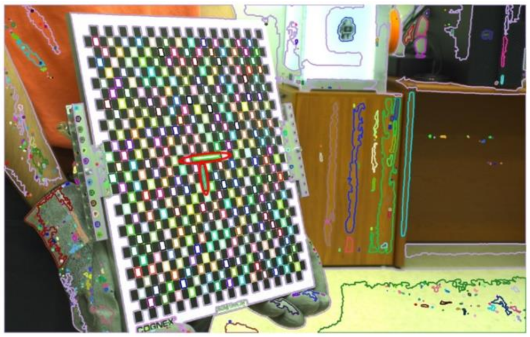


Figure 12. Detected contours from the stereo calibration image

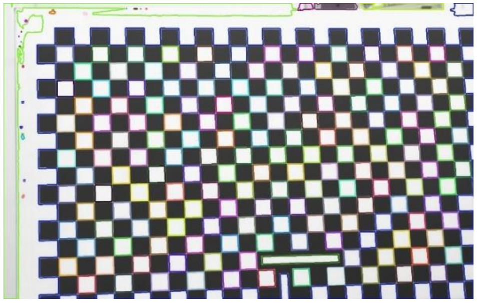


Figure 13. Detected contours from the single camera calibration image

Significant improvements to the algorithms are developed for corner auto-detection in stereo calibrations. One of them is the square and rectangle detection algorithm. Because all the squares and rectangles do not appear perfect when the board looks smaller, the new algorithm relaxes the detection criteria and implements a new concept of "two-point contour" to better detect the rectangle. The iteration procedure for corner searching is also modified in the final step to remove the remaining false detections. Fig. 14 shows the detected corners using the improved algorithms. Results show significant improvements compared with the results shown in Fig. 12.

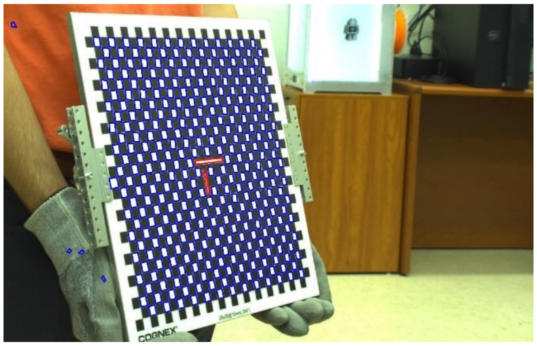


Figure 14. Improved detected corners from the stereo calibration image

Since no need to calibrate distortion, the checkerboard is placed parallel to the two camera's centerline. About 3-6 positions are needed to be placed from the near to far range.

Even with the improvements, some extreme cases may fail auto-detection when images are very blurry. To address this issue, an algorithm is developed to detect the condition and remove the blurry images automatically. These low-quality images are highlighted to users. They can be used to train users in understanding the image quality and improving skills when taking measurements to provide better quality calibration images.

The calibration final results are saved in both xml and yaml format for easy export/import between Python, C++, or Matlab code. The saved information includes: left camera matrix, right camera matrix, left camera distortion coefficients, right camera distortion coefficients, 3x3 rotation matrix between two cameras, 1x3 translation matrix between two cameras, fundamental matrix, essential matrix, and the projection error.

In summary, automatic camera calibration methods are developed using a calibration board with markers. The auto-calibration procedure includes both single camera calibration and stereo camera calibration. Advanced algorithms are developed to improve calibration accuracy via improving corner detection in sub-pixel accuracy. Improvements are also made for algorithm robustness to enhance the handling of the complex background. This auto-calibration procedure eliminates the need for users to click hundreds of times and significantly reduces the problems caused by human errors.

CONCLUSION

Manufacturers are facing challenges in the robot's accuracy assessment and accuracy improvement. NIST's development of the vision-based 6-D sensing system enables the capture of the robot's dynamic movements in high accuracy. This sensing technology provides a solution to allow the robot system's health monitoring and assessment. As the foundational supporting technology, a high-efficiency and high-accuracy vision-based system calibration procedure is critical for real applications. The developed sub-pixel corner extraction and iteration improvements enable the high accuracy of sensing system calibration. The automatic procedure and the separation of the distortion calibration with stereo calibration significantly reduce the time for on-site adjustment. The verification method enables the quick check of the system accuracy. The 6-D sensing system was used on a Universal Robot (UR5) to monitor and assess the robot accuracy degradation with different payloads, speeds, and temperatures. The test data set was published in [20]. NIST is actively seeking to develop additional industrial use cases using the 6-D sensing system 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] E. Khalastchi and M. Kalech, "Fault Detection and Diagnosis in Multi-Robot Systems: A Survey," *Sensors (Basel),* vol. 19, no. 18, p. 4019, 2019, doi: 10.3390/s19184019.

[2] M. A. Costa, B. Wullt, M. Norrlöf, and S. Gunnarsson, "Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting," *Measurement :Journal of the International Measurement Confederation,* vol. 146, pp. 425-436, 2019, doi: 10.1016/j.measurement.2019.06.039.

[3] S. Fan, L. M. Zhang, J. Wang, Y. F. Wang, Q. S. Zhang, and H. Zhao, "Vision-Based Fault Classification for Monitoring Industrial Robot," *2018, Conference Proceedings: Technical Committee on Control Theory, Chinese Association of Automation*, pp. 5889-5894, doi: 10.23919/ChiCC.2018.8483793.

[4] E. Sita, T. Thomessen, A. G. Pipe, M. Studley, and F. Dailami, "Usability Study of a Robot Companion for Monitoring Industrial Processes," 2020, *Conference Proceedings: IEEE, p*p. 37-42, doi: 10.1109/ACIRS49895.2020.9162607.

[5] R. N. A. Algburi and H. Gao, "Health Assessment and Fault Detection System for an Industrial Robot Using the Rotary Encoder Signal," *Energies (Basel),* vol. 12, no. 14, p. 2816, 2019, doi: 10.3390/en12142816.

[6] S. K. Mustafa, P. Y. Tao, G. Yang, and I. Chen, "A geometrical approach for online error compensation of industrial manipulators," in *2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 6-9 July 2010 2010, pp. 738-743, doi: 10.1109/AIM.2010.5695784.

[7] B. C. Jiang, R. Duraisamy, G. Wiens, and J. T. Black, "Robot metrology using two kinds of measurement equipment," *Journal of Intelligent Manufacturing,* vol. 8, no. 2, pp. 137-146, 1997/03/01 1997, doi: 10.1023/A:1018508805175.

[8] J. H. Jang, S. H. Kim, and Y. K. Kwak, "Calibration of geometric and non-geometric errors of an industrial robot," *ROBOTICA,* vol. 19, pp. 311-321, 2001.

[9] S. Wang, S. Liu, and Q. Mao, "A CMM-Based Method of Control Point Position Calibration for Light Pen Coordinate Measuring System," *Sensors,* vol. 20, no. 19, p. 5592, 2020, doi: 10.3390/s20195592.

[10] A. Wiles, D. Thompson, and D. Frantz, "Accuracy assessment and interpretation for optical tracking systems" *SPIE 2004*. Proceedings vol. 5367, 2004, doi:10.1117/12.536128.

[11] T. Sun, Y. Zhai, Y. Song, and J. Zhang, "Kinematic calibration of a 3-DoF rotational parallel manipulator using laser tracker," *Robotics and computer-integrated manufacturing,* vol. 41, pp. 78-91, 2016, doi: 10.1016/j.rcim.2016.02.008.

[12] G. Qiao and B. A. Weiss, "Industrial Robot Accuracy Degradation Monitoring and Quick Health Assessment," *Journal of Manufacturing Science and Engineering,* vol. 141, no. 7, 2019, doi: 10.1115/1.4043649.

[13] G. Qiao, C. Schlenoff, and B. A. Weiss, "Quick positional health assessment for industrial robot prognostics and health management (PHM)," 2017, *Conference Proceedings: IEEE*, pp. 1815-1820, doi: 10.1109/ICRA.2017.7989214.

[14] O. Nina, B. Morse, and W. Barrett, "A recursive Otsu thresholding method for scanned document binarization," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, 5-7 Jan. 2011 2011, pp. 307-314, doi: 10.1109/WACV.2011.5711519.

[15] T. Panchal, H. Patel, and A. Panchal, "License Plate Detection Using Harris Corner and Character Segmentation by Integrated Approach from an Image," *Procedia Computer Science,* vol. 79, pp. 419-425, 2016/01/01/ 2016, doi: https://doi.org/10.1016/j.procs.2016.03.054.

[16] A. Datta, J.-S. Kim, and T. Kanade, "Accurate camera calibration using iterative refinement of control points,"

2009, *Conference Proceedings: IEEE*, pp. 1201-1208, doi: 10.1109/ICCVW.2009.5457474.

[17] K. H. Strobl and G. Hirzinger, "More accurate pinhole camera calibration with imperfect planar target," 2011, *Conference Proceedings: IEE*E, pp. 1068-1075, doi: 10.1109/ICCVW.2011.6130369.

[18] C. Galamhos, J. Matas, and J. Kittler, "Progressive probabilistic Hough transform for line detection," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 23-25 June 1999 1999, vol. 1, pp. 554-560 Vol. 1, doi: 10.1109/CVPR.1999.786993.

[19] A. Trujillo-Pino, K. Krissian, M. Alemán-Flores, and D. Santana-Cedrés, "Accurate subpixel edge location based on partial area effect," *Image and vision computing,* vol. 31, no. 1, pp. 72-90, 2013, doi: 10.1016/j.imavis.2012.10.005.

[20] G. Qiao, "Degradation measurement of robot arm position accuracy," 2019, doi: https://www.nist.gov/el/intelligent-systems-division-73500/degradation-measurement-robot-arm-position-accuracy.

The 16th Conference of the International Society of Indoor Air Quality & Climate (Indoor Air 2020)
COEX, Seoul, Korea | July 20 - 24, 2020

Paper ID ABS-0446 ISBN: …………………….

Development and Application of an Indoor Carbon Dioxide Metric of Ventilation

Andrew Persily[1,*], Brian Polidoro[1]

[1] National Institute of Standards and Technology, Gaithersburg, USA

[*] *Corresponding email: andyp@nist.gov*

SUMMARY

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. After many unsuccessful efforts to dissuade practitioners and researchers from such uses and abuses of indoor CO_2, an approach has been developed to estimate CO_2 concentrations to serve as indicators of outdoor air ventilation rates per person, though not overall IAQ. Rather than a single concentration for all spaces and occupancies, a space-specific CO_2 metric of ventilation is proposed that is based primarily on recommended or required ventilation rates per person and occupant characteristics. This paper describes the approach, with sample calculations discussed for several commercial, institutional and residential building occupancies. An online calculator to perform these calculations is also described.

KEYWORDS

carbon dioxide; indoor air quality; metrics; ventilation

1 INTRODUCTION

Indoor air quality (IAQ) is characterized by the presence and concentrations of chemical and physical substances in air that impact occupant health, comfort and productivity. The number of airborne contaminants in most indoor environments is large, and their impacts on building occupants are known for only a limited number of substances. The wide variation in contaminants 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 fully capture the health and comfort impacts of IAQ nor have any been accepted in the field (Jackson et al., 2011; Hollick and Sangiovanni, 2000; Moschandreas et al., 2005; Teichman et al. 2015).

For many years, some have promoted indoor CO_2 concentrations as a metric of IAQ and ventilation, in many cases without a clear explanation of the application and associated limitations (Persily 1997). Practitioners and researchers frequently use 1800 mg/m^3 (roughly 1000 ppm$_v$) as such a metric, often erroneously basing it on ASHRAE Standard 62.1 (ASHRAE 2019a). However, that standard has not contained an indoor CO_2 limit for almost 30 years (Persily 2015). There have been many papers and presentations that have attempted to clarify the significance of indoor CO_2 concentrations, with some advocating that they not be used in IAQ and ventilation evaluations. However, calls to stop poorly informed applications of indoor CO_2 are not succeeding. While efforts to educate designers, practitioners and others about the application of indoor CO_2 need to continue, this paper presents an approach for using indoor CO_2 concentrations as a metric of ventilation rate per person that considers the space-specific parameters that determine indoor CO_2 levels and describes an on-line calculator to estimate indoor CO_2 concentrations for use with this approach (Persily, 2018; Persily and Polidoro, 2019). Note that this approach does not involve using CO_2 concentrations as an IAQ metric, but rather to evaluate per person ventilation rates using indoor CO_2 as a tracer gas. While indoor CO_2 concentrations are linked to the perception of odors associated with the byproducts of human metabolism and other contaminants associated with the number of occupants, there are many other important indoor air contaminants that are not associated with the number of occupants, and CO_2 is not a good indicator of those contaminants.

2 CO_2-BASED VENTILATION METRIC

While a CO_2 concentration metric that characterizes IAQ would be attractive, such a metric is not possible, in part because there are many indoor air contaminants with significant health and comfort impacts and whose concentrations are unrelated to CO_2 concentrations. Previous work describes the use of CO_2 as an indicator or metric of outdoor air ventilation rates per person (Persily, 2018; Persily and Polidoro, 2019). As discussed in those papers, indoor CO_2 concentrations depend primarily on the rate *G* at which occupants generate CO_2, the outdoor air ventilation rate of the space, the time since occupancy began, and the outdoor CO_2 concentration. Note that outdoor air ventilation refers to the total rate at which outdoor air enters a building or space, including mechanical and natural ventilation as well as infiltration. The cited papers describe the single-zone mass balance theory for calculating indoor CO_2 concentrations, noting that the indoor concentration will only achieve steady-state if conditions, specifically ventilation rate and *G* are constant for a long enough period of time. Those papers stress the assumptions on which the single-zone mass balance is based, including that the concentration in the space can be characterized by a single value, which some describe as the well-mixed assumption. Also, a constant *G* requires that occupancy and activity remain constant, but in many spaces both will be too short or too variable for steady-state to occur. 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 the inverse of the outdoor air change rate; under constant occupancy and activity the indoor concentration will be about 95 % of steady-state after three time constants.

2.1 Commercial and Institutional Buildings

In the previous work on this CO_2 ventilation metric (Persily 2018), several space types were selected from the commercial/institutional building space types or “Occupancy Categories” in Table 6-1 of ASHRAE Standard 62.1 (ASHRAE 2019a). For these spaces, as 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 based on the judgement of the authors. The assumed occupant densities, occupant characteristics and CO_2 generation rates are described in (Persily 2018).

Table 1. Calculated CO_2 concentrations in commercial and institutional occupancies

			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	146	1170	1450

* Time to achieve 95 % of steady-state CO_2 concentration, i.e., three time constants

** To convert these concentrations too ppm_v, divide these values by 1.8.

For most of the spaces, the time to steady-state is less than 1.5 h. In those cases, the three calculated concentrations are generally within 100 mg/m^3, making timing of a measurement for comparison to a ventilation metric less critical, though such a measurement still needs to take place about an hour after occupancy starts. For spaces with longer times to steady-state, the three CO_2 concentrations cover a broader range. For those spaces, the CO_2 concentration after 1 h will be more sensitive to measurement timing than the t_{metric} or steady-state values. The office space takes almost 6 h to reach steady-state due largely to its low occupant density and low air change rate. As a result, the three concentrations are quite different. It's unlikely for many office spaces to be at full occupancy for 6 h given lunch breaks and other common events; therefore, the t_{metric} value of 2 h and the corresponding concentration of about 600 mg/m^3 are more relevant. These same calculations are presented in Persily (2018) for ventilation rates 25 % lower than those assumed here, given the desire for a CO_2-based ventilation metric to capture situations in which ventilation rates are lower than intended. Based on this previous work, including the desire for a CO_2 metric to identify ventilation deficiencies and to be less sensitive to the timing of the concentration measurement, Table 2 contains potential CO_2 metric values for these spaces along with the corresponding measurement time. 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 inputs would possibly yield other potential metric values.

Table 2. Potential CO_2 concentration metrics of ventilation

Space type	CO_2 concentration metric, above outdoors mg/m^3 (µL/L*)	Corresponding time, h after full occupancy
Classroom (5 to 8 y)	1000 (550)	2
Classroom (>9 y)	1500 (850)	2
Lecture classroom	2000 (1100)	1
Restaurant	2000 (1100)	2
Conference room	2000 (1100)	1
Hotel/motel bedroom	1000 (550)	6
Office space	600 (350)	2
Auditorium	3000 (1700)	1
Lobby	4500 (2500)	1
Retail/Sales	1500 (850)	2

* µL/L is equivalent to ppm_v; values in parentheses are approximate conversions

2.2 Residential Buildings

Extending this approach to residential spaces is challenging given the wide variations in dwelling and family size and in occupant characteristics, as well as the often unpredictable durations of occupancy relative to many commercial and institutional spaces. However, the hours associated with sleep provide longer periods of constant occupancy to support analyses in bedrooms. The approach for residential buildings in Persily and Polidoro (2019) is to again to use single zone, steady-state mass balance analysis to calculate the CO_2 concentrations for a given space based on assumptions about CO_2 generation rate and ventilation rate. In order to explore these dependencies for residential spaces, indoor CO_2 concentrations were calculated for three 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 used to calculate the CO_2 generation rate for each person during non-sleep hours when occupants are assumed to be more active, and when occupants are sleeping.

For each occupancy, CO_2 concentrations were calculated for several different ventilation scenarios. Whole house concentrations were calculated using rates based on ASHRAE Standard 62.2 and a fixed rate of 0.5 h^{-1}. Bedroom concentrations were calculated based on Standards 62.2, 0.5 h^{-1} and 10 L/s per person. For the bedroom cases, two idealized outdoor 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 then 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. In the other case, Uniform Distribution, the total ventilation 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. For the case with bedroom ventilation rates of 10 L/s per person, 10 L/s of outdoor air is supplied for each person in each bedroom, with that rate based on recommendations in CEN (2007a) and (2009).

Persily (2019) presents calculated CO_2 concentrations for these different residential occupancies and ventilation scenarios. In contrast to the commercial and institutional occupancies discussed earlier, the variation in space size, occupancy and ventilation in these residential cases makes it difficult to generalize these results. Instead, the house, occupancy and air distribution approach all need to be accounted for in developing a metric or reference point for evaluating the adequacy of the outdoor air ventilation rate relative to a target value. The online tool discussed in the next section was developed to implement these concepts.

3 ON-LINE CO_2 METRIC CALCULATOR

To support application of the proposed CO_2 metric concept for assessing ventilation rates, an online tool has been developed (https://pages.nist.gov/CONTAM-apps/webapps/CO2Tool/#/). 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 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 for all indoor spaces.

When using the tool, the user first selects whether they wish to analyze a commercial/institutional building or a residential building and 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 62.1-2019, and to use the default values in that standard for outdoor ventilation requirements and occupant density. 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. However, all of these inputs can be modified by the user. For residential buildings, the user selects a whole building or bedroom analysis. If whole building, the user can use of a ventilation requirement based on Standard 62.2-2019 or enter a whole building air change rate in h^{-1}. If instead they are performing a bedroom analysis, they need to select a whole building ventilation requirement, e.g. using 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. For both commercial/institutional or residential buildings, an Alternate ventilation rate can be input to compare the results to those obtained with a Primary ventilation rate.

Once the user has completed the inputs, a Results screen 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 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. 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 building or space under consideration, and the calculated value at that time is then used for the comparison. If constant occupancy does not last that long, the t_{metric} value in the calculator can be modified.

For bedrooms, the calculated CO_2 concentrations are also intended to be compared to measured values. 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. A measured value that is higher than the calculated is an indication that the actual ventilation rate is below the assumed or desired rate. Note that this comparison neglects the impact of interzone CO_2 transport on the bedroom concentration.

4 CONCLUSIONS

This paper presents an 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. It is not intended to serve as an IAQ metric, only to assess ventilation rates. Application of this approach requires one to report information on the space being considered and its occupancy, the time at which full occupancy starts, time of the CO_2 concentration measurement, and the 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 62.1, Standard 62.2 or another ventilation requirement of interest. As additional analyses are performed and the concept is discussed with practitioners and researchers, it is anticipated that the approach will evolve and become more well-defined and useful. An online calculator has been developed to allow users to exercise this approach. 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).

Note that there are several important limitations with this approach. Most importantly, it only provides a metric of ventilation rates per person; it will not characterize overall IAQ or the concentrations of other indoor air contaminants that impact occupant health, comfort and productivity. Also, it has only been applied to a limited number of cases. Additional analyses with this approach are needed to better understand its application and usefulness for different building types and occupants. The single-zone analysis approach does not account for air distribution impacts on spatial variations in CO_2 concentrations nor does it account for transient effects associated with changes in occupant activities.

7 REFERENCES

ASHRAE. 2019a. ANSI/ASHRAE Standard 62.1-2019 Ventilation for Acceptable Indoor Air Quality, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

ASHRAE. 2019b. ANSI/ASHRAE Standard 62.2-2019 Ventilation and Acceptable Indoor Air Quality in Low-Rise Residential Buildings, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

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. 2009. Ventilation for buildings - Determining performance criteria for residential ventilation systems, Brussels, European Committee for Standardization.

Hollick HH and Sangiovanni JJ. 2000. A Proposed Indoor Air Quality Metric for Estimation of the Combined Effects of Gaseous Contaminants on Human Health and Comfort, In: Nagda NL (ed) Air Quality and Comfort in Airliner Cabins, ASTM STP 1393, West Conshohocken, PA, American Society for Testing and Materials, 76-98.

Jackson MC, Penn RL, Aldred JR, Zeliger HI, Cude GE, Neace LM, Kuhs JF and Corsi RL. 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.

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 AK. 1997. Evaluating Building IAQ and Ventilation with Indoor Carbon Dioxide. *ASHRAE* Transactions, 103 (2), 193-204.

Persily A. 2015. Challenges in developing ventilation and indoor air quality standards: The story of ASHRAE Standard 62. Building and Environment, 91, 61-69.

Persily A. 2018. Development of an Indoor Carbon Dioxide Metric, 39th AIVC Conference, Antibes Juan-les-Pins, France, 791-800.

Persily A, and Polidoro B. 2019. Residential Application of an Indoor Carbon Dioxide Metric, 40th AIVC Conference, Ghent, Belgium, 995-1007.

Teichman K, Howard-Reed,C, Persily A and Emmerich S. 2015. Characterizing Indoor Air Quality Performance Using a Graphical Approach, National Institutte of Standards and Technology.

Advancing the Accuracy of Computational Models for Double-Sided Incremental Forming

Newell Moser [a, *], Dohyun Leem [b], Shuheng Liao [b], Kornel Ehmann [b], Jian Cao [b]

[a] National Institute of Standards and Technology, Applied Chemicals and Materials Division,
325 Broadway, Boulder, CO, 80305, USA
[b] Northwestern University, Department of Mechanical Engineering, 2145 Sheridan Rd,
Evanston, IL, 60201, USA
***** Corresponding author, Phone: +1 (303) 497-5711, Email: newell.moser@nist.gov

Newell Moser: newell.moser@nist.gov
Dohyun Leem: dohyunleem2021@u.northwestern.edu
Shuheng Liao: shuhengliao2023@u.northwestern.edu
Kornel Ehmann: k-ehmann@northwestern.edu
Jian Cao: jcao@northwestern.edu

Abstract

Double-Sided Incremental Forming (DSIF) is a rapid-prototyping manufacturing process for metal forming that, for low-volume production, is competitively energy-efficient. However, controlling the DSIF process in terms of accuracy and formability is an ongoing challenge. These control challenges arise due to a lack of understanding of the underlying deformation mechanisms in DSIF, which finite element simulations can help to unravel. However, DSIF pushes the limits of modern finite element formulations due to true strains that approach one, finite rotations, nonlinear contact, and triaxial stress states that range across multiple length scales. To confidently develop a finite element model of DSIF, an extensive verification process must be considered, which is the objective of this study. In this work, different finite element types and varying amounts of artificial acceleration are investigated, and recommendations based on efficiency and accuracy are summarized. A simplified, axisymmetric geometry was considered to reduce simulation time. For this geometry, accelerating the explicit finite element simulation by a mass factor of 10^5 or greater affected the stress triaxiality in the sheet by as much as 40% in some locations with respect to the quasi-static case. Additionally, the ratio of the kinetic energy to internal energy of the sheet was not a reliable indicator of whether a DSIF simulation is approximately quasi-static.

Keywords

Metal forming, Simulation, Finite element, Verification, Mass scaling

1. Introduction and Motivation

Double-Sided Incremental Forming (DSIF) is a flexible metal forming process that uses two numerically controlled tools – one above and one below the sheet – that move along a predefined toolpath [1]. In addition to flexibility, parts formed by DSIF demonstrate a marked improvement in formability relative to conventional forming processes like stamping [2]. More specifically, failure strains in DSIF surpass those observed in stamping, which are usually defined by the necking limit and are characterized by forming limit diagrams.

Although DSIF has great potential towards the rapid production of freeform sheet metal parts, the use of a secondary tool also contributes to additional challenges that can affect process control. Similar to Single-Point Incremental Forming (SPIF), the principal technical challenges in DSIF are related to process control: geometric accuracy, surface quality, process throughput, and excessive thinning or fracture [3]. In order to improve process control, there is a need to develop accurate, efficient models of DSIF in order to better understand the mechanics of the process, as well as to quantify values which cannot be easily measured (e.g., residual stresses).

There are numerous challenges associated with simulating DSIF. Even forming a simple geometry, like a funnel, involves large-strain deformations, finite rotations, cyclic loading-unloading histories, nonlinear material behavior, contact-impact interfaces, and a range of triaxial stress states. To overcome these challenging nonlinearities, state-of-the-art techniques within the finite element method are commonly used [4–7].

Without any artificial acceleration, a finite element simulation of DSIF could readily require *months* to finish [8]. One strategy to speed up a simulation is to choose a simpler geometry, like an axisymmetric part, and develop a reduced model [7]. Though, for large finite element problems, explicit (transient) integrators with artificial acceleration tend to scale better than implicit (equilibrium) integrators, particularly if strong nonlinearities are present like contact interfaces. For a more detailed discussion on explicit and implicit finite element methods, the interested reader is referred to Belytschko et al. [9] and Wriggers [10]. Since full-scale simulations of DSIF often involve a large number of finite elements (greater than 10^6 degrees of freedom), explicit methods are a popular choice to model the process [6, 11].

To artificially accelerate an explicit finite element model, either mass scaling or velocity scaling can be used. By increasing the mass in the model, the stable time step associated with explicit simulations increases, thereby necessitating fewer time steps to complete the simulation. Researchers simulating incremental sheet forming commonly use a mass scaling factor of 10^5, or more [6, 11, 12]. On the other hand, researchers must be careful not to use too much artificial acceleration. Otherwise, the model will no longer be a fair approximation of the real process, which is usually quasi-static depending on the tool speed. However, for a given DSIF model, it is currently unclear what a safe upper limit is in terms of mass scaling.

In this study, we develop a simplified finite element model of DSIF in order to verify the relative effects of varying key parameters in the model. By changing the parameters in the model, like mass scaling, the predicted stresses and strains will be shown to significantly change relative to the baseline quasi-static simulation. As a result, recommended values and methods related to mass scaling and finite element type are given and discussed.

2. Verification Model of DSIF

In order to vary a wide range of finite element parameters within a practical time span, a simplified DSIF model was created which is still capable of capturing the predominant forming mechanics of DSIF. As shown in Figure 1, an axisymmetric, partially-preformed part was chosen, and only half of the model was meshed by using a symmetry boundary condition. In DSIF, shear distortion occurs tangential to the toolpath, so a 10 mm extension was created to distance any erroneous effects of the symmetry boundary condition from the conical region of interest. Although the symmetric boundary condition is not strictly true in DSIF, it is assumed to be a fair assumption since the region of interest is far away from the boundary and the dominant mechanics of DSIF are highly localized. To clarify, the region of interest is within the center of the toolpath's arc just as the tool crosses the X-axis.

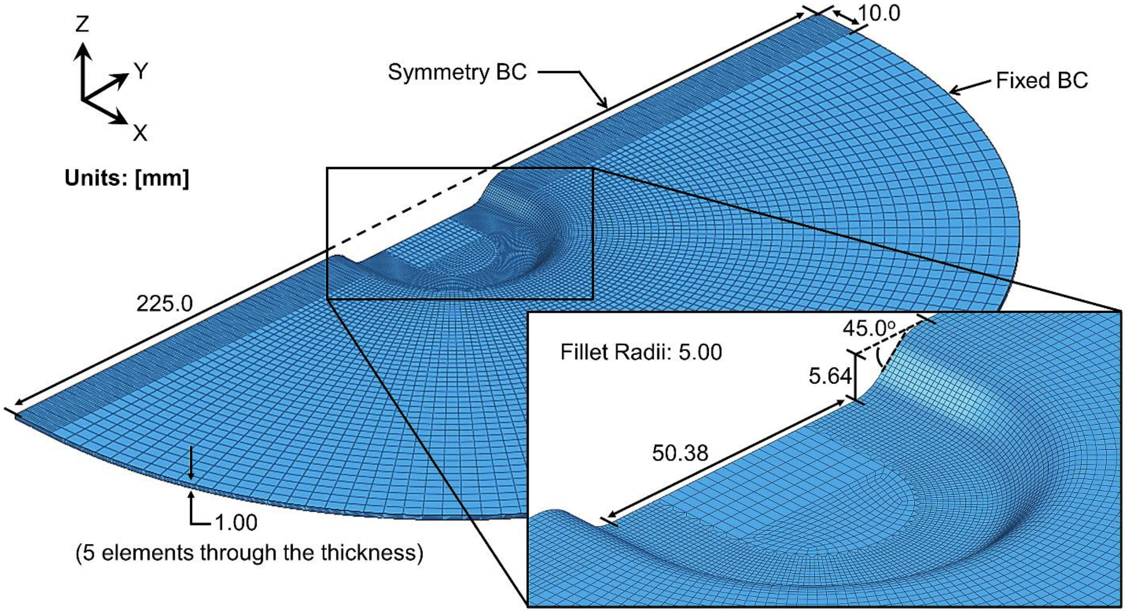


Figure 1. Shown is the finite element mesh used for the verification DSIF model.

For each revolution in the spiral toolpath, a "jump" has been added so that the tools lift-off and then regain contact while traversing along the symmetry boundary condition. By reducing the domain size and performing only five toolpath revolutions for a total length of 731 mm, this simplified model serves as an efficient testbed for comparing finite element procedures in DSIF. However, the proposed verification model is only suggested for investigating *relative* comparisons between simulations.

A total of 55,096 reduced-integration, linear brick elements with stiffness-based hourglass controls [13] were used to model the sheet in Figure 1, while the hemispherical tools – 10 mm in diameter – were modeled with rigid shell elements. Although reduced-integration elements were ultimately chosen, other element formulations were trialed, which will be discussed in the next section. In the forming region, elements were approximately 0.5 mm × 0.5 mm × 0.2 mm in size. A node-to-surface contact constraint was enforced using the penalty formulation, and friction was neglected. The material of the sheet was modeled after aluminum alloy 5754-O using J_2 flow theory, also commonly referred to as von Mises plasticity with isotropic hardening. A Voce strain hardening model was used within the material model, based on the same parameters used by Moser et al. [4].

All of the verification simulations were performed using Abaqus FEA[1] (version 2017). Since this verification model is simplified and less costly in computational resources than a full-scale model of DSIF, it was reasonable to solve this representative DSIF problem using an implicit (equilibrium) scheme. This static case was used to establish a baseline. Then, the verification model was solved using an explicit (transient) scheme with successive increases in artificial acceleration.

3. Finite Element Type

Different hexahedral element technologies were trialed using the verification model shown in Figure 1 while assuming static conditions. Shell element formulations were not considered in this work due to the underlying assumptions in shell elements related to stresses and strains in the thickness direction, which can be inaccurate for double-sided contact [14]. A summary of the key findings is given in Table 1. Abaqus nomenclature is used as shorthand for the different types of elements: C3D20R is a 20-node, second-order serendipity element that uses reduced integration and stiffness-based hourglass controls; C3D8 is an 8-node, first-order element based on a selectively reduced integration scheme; C3D8I is an 8-node, first order, fully-integrated element that is enhanced by incompatible modes; and C3D8R is an 8-node, reduced-integration element that utilizes stiffness-based hourglass controls. In an effort to be comprehensive, shear locking, volumetric locking, and hourglass controls are briefly summarized next; more details are discussed by Belytschko et al. [9].

Shear locking occurs in first-order, fully-integrated elements that are subjected to bending. The numerical formulation of these elements gives rise to shear strains that do not really exist – the so-called parasitic shear. Therefore, these elements are too stiff in bending. Volumetric locking (sometimes termed pressure locking) occurs in fully-integrated elements when the material behavior is incompressible, such as von Mises materials. Spurious pressure-stresses develop at the integration points, causing an element to behave too stiffly for deformations that should cause no volume changes (see Figure 2). Reduced-integration, or under-integrated, refers to using fewer Gaussian integration points than are necessary to integrate the element matrices exactly for an undistorted element. A linear (i.e., first-order) element with reduced-integration implies that a single integration point is used, which is located at the element center.

[1] Certain commercial equipment, software and/or materials are identified in this paper in order to adequately specify the experimental or simulation procedure. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment and/or materials used are necessarily the best available for the purpose.

Table 1. Results of trialing different solid continuum elements in the verification DSIF model. Wall time is defined as the elapsed "wall-clock" time required to complete the simulation. HG is short for hourglass, and SR for selectively reduced.

Type of Hexahedral Solid Continuum Element	Shear Locking	Volumetric Locking	Wall Time × Number of Computer Cores (CPU · hr)
C3D20R (w/ HG Controls)	Minimal	Severe	(256.3 hr) × (72 CPU) = 18,454
C3D8 (SR Integration)	Severe	None	(52.5 hr) × (24 CPU) = 1,260
C3D8I (Incompatible Modes)	Severe	Moderate	(86.2 hr) × (24 CPU) = 2,069
C3D8R (w/ HG Controls)	None	None	(60.2 hr) × (24 CPU) = 1,446

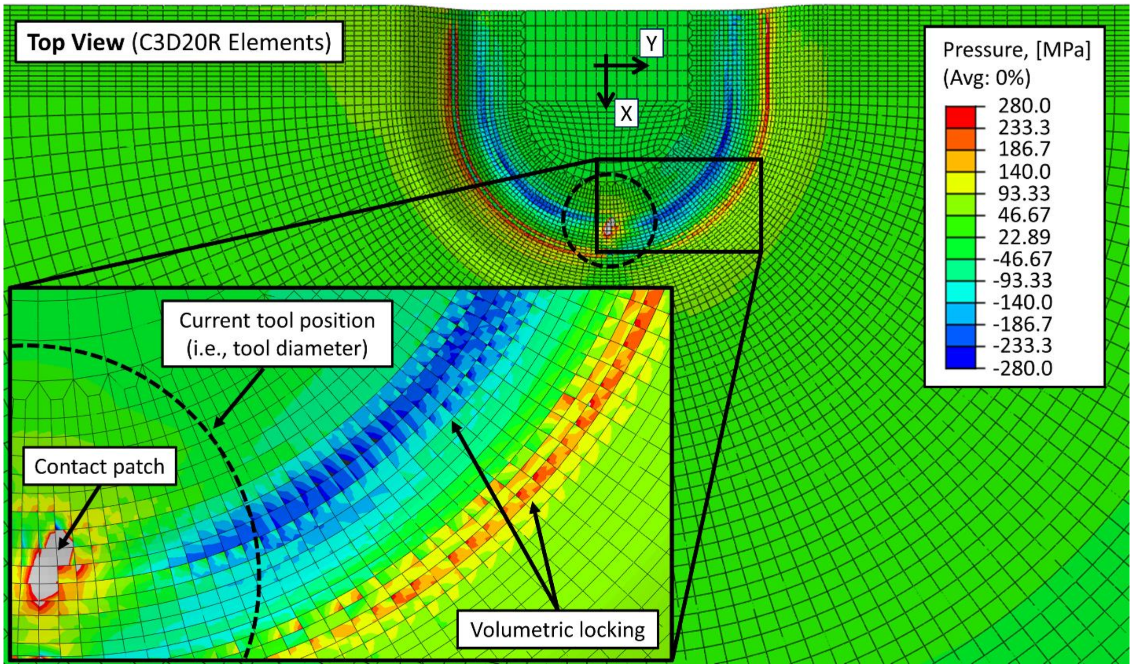


Figure 2. An example of volumetric locking in the pressure (i.e., negative mean stress) field when using quadratic C3D20R elements. Notice the strong discontinuities across element edges. The top and bottom tools (currently hidden) are traversing counterclockwise.

Referring to Table 1, fully-integrated hexahedral elements (C3D9I) fall victim to volumetric locking and shear locking. However, a reduced-integration (or selectively reduced-integration) scheme for the element stiffness matrix not only greatly reduces running time, but these formulations can also alleviate volumetric locking and shear locking. Though, under-integrated elements can result in spurious zero-energy modes of deformation, which are commonly termed "hourglass" modes. Thus, hourglass controls must be employed to artificially stiffen the response of the element against these zero-energy deformation modes. Moreover, hexahedral elements usually provide a solution of equivalent accuracy at less cost compared to tetrahedral elements [9]. However, hexahedral elements are more sensitive to severe mesh distortions, so a structured mesh should be used wherever possible. For these reasons, the C3D8R element with a structured mesh was chosen for this simplified DSIF model.

4. Effects of Mass Scaling

An explicit finite element simulation can be sped up by artificially increasing the mass or scaling the velocity of moving entities. More specifically, increasing the velocity by a factor of k reduces the simulation time by a factor of k, but also artificially increases the kinetic energy by k^2. Increasing the mass (or density) by a factor of k reduces the simulation time by $\sqrt{k}$, but artificially increases the kinetic energy by k. For a given artificial increase in kinetic energy, the effects of accelerating the simplified DSIF model will be similar whether it is achieved through mass or velocity scaling. So, only mass scaling is discussed here.

Referring to Table 2, four cases were considered in evaluating the effects of mass scaling: 1) the Implicit Static case based on solving the static equations using an implicit load-stepping algorithm, 2) the Quasi-Static case based on no mass scaling, 3) the Mass-1 case based on a mass-scaling factor of 1.0×10^3, and 4) the Mass-2 case based on a mass-scaling factor of 1.0×10^5. The tool velocity was set to be constant for all cases involving mass scaling.

Table 2. Mass scaling factors considered in the verification DSIF model. Except for the Implicit Static model, all models utilize an explicit time-stepping algorithm.

Nickname	**Implicit Static**	**Quasi-Static**	**Mass-1**	**Mass-2**
Mass Scaling Factor	1.0	1.0	1.0×10^3	1.0×10^5
Tool Velocity	~	250 mm/s	250 mm/s	250 mm/s
Wall Time (72 cores)	20.1 hr	256.5 hr	8.7 hr	0.9 hr

The stress triaxiality through the thickness during the final toolpath revolution is shown in Figure 3 for each simulation case. Note, the stress triaxiality is defined as the hydrostatic stress divided by the von Mises equivalent stress. With mass scaling, the stress triaxiality lessens in magnitude (becomes more compressive) near regions of contact. More specifically, adding mass by a factor of 10^5 changed the stress triaxiality by 41% near the tip of the top tool, and by 23% near the bottom tool. Mass-1 utilized a moderate amount of mass scaling, and the trends and magnitudes in the stress triaxiality are not significantly altered relative to the Quasi-Static case. Although not shown here, increasing amounts of mass scaling had less of an effect on kinematic quantities – like the equivalent plastic strain – compared to kinetic quantities, like the stress triaxiality and the Lode angle parameter.

Another interesting observation to take note of in Figure 3 is that the Quasi-Static case, based on explicit time-stepping, predicts higher stress triaxialities compared to the Implicit Static case; ideally, they should closely match. We believe these differences are largely due to the numerical implementations of the explicit and implicit procedures, and not because of the physics of the problem at hand. The largest differences in the predictions occurred near regions of contact, which supports the assumption that these effects could be due to the difference in the implementation of the implicit and explicit contact algorithms. Both the implicit scheme (dependent on the iterative Newton's method) and the explicit scheme (based on the central difference method) are second-order accurate in time regarding the truncation error [9]. However, the iterative Newton scheme inherent to implicit methods requires that the energy and displacement residual errors be within a specified tolerance in order to meet convergence requirements. Explicit methods, on the other hand, do not directly enforce the energy balance; they take a small step forward in time without the need for iterations. Furthermore, explicit methods rely on lumping (i.e., diagonalizing) the mass matrix rather than using the consistent mass matrix, and the lumping procedure is generally performed in an *ad hoc* fashion [15]. Although we can speculate why these differences in the stress triaxialities are appearing, we cannot be certain which is more accurate since this study is limited to relative comparisons. Still, it is interesting to note that, with everything else being equal, the implicit and explicit schemes predicted different magnitudes in the stress triaxialities; though, they do predict similar trends.

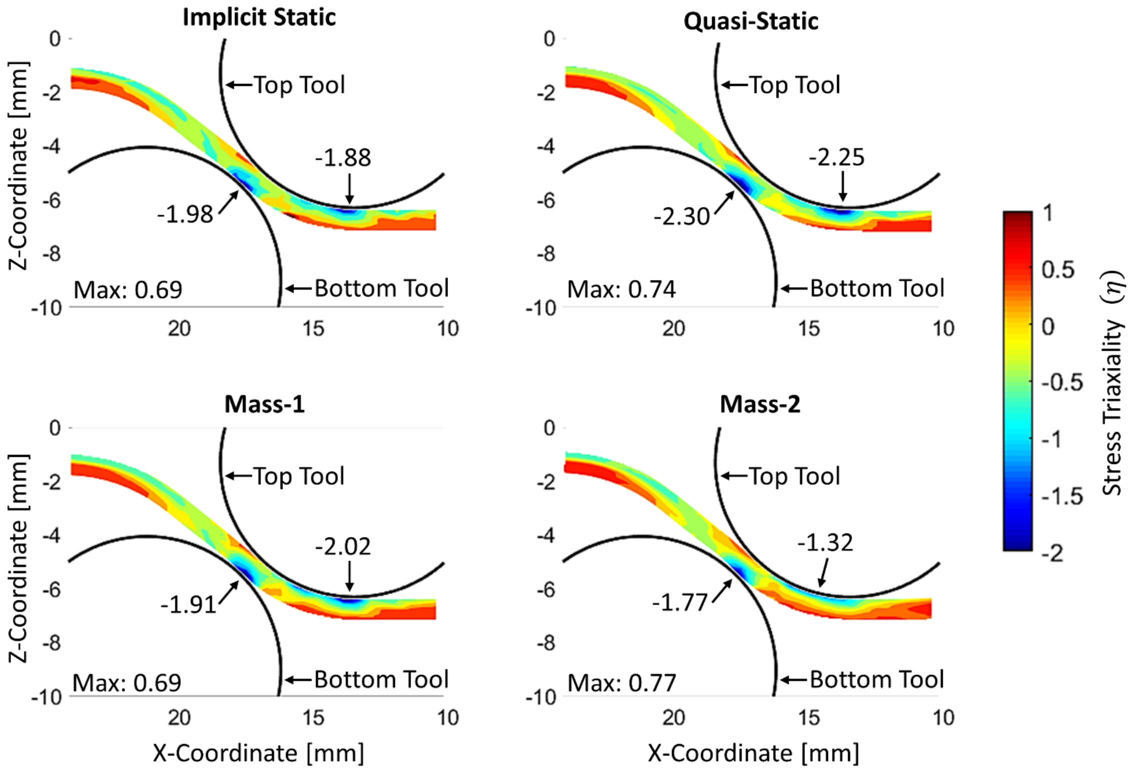


Figure 3. The relative effects of mass scaling on the stress triaxiality. Shown in each case is the cross-sectional side view of the sheet while the tools pass over the X-axis (see Figure 2). The black partial circles represent the perimeter of the tools at this time in the simulation. The minimum (i.e., most compressive) stress triaxialities are pointed out near the regions of contact.

The average forming forces and their corresponding standard deviations are given in Table 3. Although the average magnitudes of the forming forces vary due to mass scaling, the results do not exhibit strong trends. This suggest that increasing the amount of artificial acceleration does not consistently change the average magnitude of the forming forces. On the other hand, increasing the amount of artificial acceleration does result in larger standard deviations of the predicted contact forces. Since the Mass-2 case is not quasi-static based on the changes in the mechanics, as shown in Figure 3, the average forming force is not a good indicator of whether a DSIF simulation is quasi-static or not.

Table 3. The Z-forces (i.e., axial forces) in the top tool are compared between the verification simulations. The average (Avg.) and standard deviation (Std. Dev.) values are calculated from the force histories throughout the final toolpath revolution.

Simulation Case	Avg. Z-Force (N)	Std. Dev. Z-Force (N)
Implicit Static	471.0	± 40.9
Quasi-Static	606.1	± 30.4
Mass-1	551.5	± 63.0
Mass-2	608.6	± 89.2

The likely reason that the average forming forces do not significantly differ in Table 3 is because the overall summation of the contact pressures do not significantly change with additional inertial effects. Even though the magnitude of the (compressive) stress triaxialities do change, the total compressive area also changes in such a way to make up for this. Then, when the contact stresses are integrated over the area to calculate the contact forces, the average resultant forces will not change significantly.

One way to determine if a metal forming simulation is quasi-static is to analyze the ratio between the kinetic energy and the internal energy of the deforming metal. In sheet metal forming, the internal energy is usually dominated by the accumulation of plastic strains. If the kinetic energy approaches the internal energy in magnitude, then the deforming sheet is no longer quasi-static. This rule-of-thumb is a reasonable assumption when the entire sheet is deforming, but it does not hold well for DSIF (see Table 4). In DSIF, the localized zones of the sheet will deform as the tools travel along the toolpath, and in effect, the (global) internal energy of the sheet continues to increase over time as new locations experience plastic straining. However, the kinetic energy remains approximately constant over time. Notice in Table 4 that the kinetic energy is at least an order of magnitude less than the internal energy for all cases. However, as stated before, the Mass-2 case is not quasi-static based on Figure 3.

Table 4. The internal energy (I.E.), kinetic energy (K.E.), standard deviation (Std. Dev.) in the kinetic energy of the deformable sheet are compared during the final toolpath revolution. All energy units are in, N·mm.

Simulation Case	Sheet I.E.	Sheet K.E.	Std. Dev. in K.E.
Implicit Static	1.11×10^4	0.00	± 0.00
Quasi-Static	1.25×10^4	0.111	± 0.088
Mass-1	1.36×10^4	22.4	± 10.6
Mass-2	1.54×10^4	366	± 118

5. Conclusions and Recommendations

The result of this verification study is a set of finite element parameters that strike a balance between computational efficiency and (relative) accuracy. The following recommendations can be considered for new, full-scale simulations of DSIF. Though, this study is not extensive; there are additional finite element parameters, like the material model, that should be validated in any new finite element model. Moreover, it is important to recall that a simplified geometry was chosen for this verification study, and consequently, we cannot guarantee that the following conclusions can be extended for all possible toolpaths in DSIF. Rather, these conclusions and recommendations should be considered as reasonable starting points, but further investigating may be necessary for specific geometries. Our conclusions and recommendations are as follows:

- As a starting point for new DSIF models, we recommend scaling the mass between a factor of 10^2 and 10^4, and in general, not above 10^5. Depending on the toolpath and mesh, adding more artificial acceleration can degrade the accuracy of the solution, specifically for kinetic terms like the stress triaxiality and Lode angle parameter.
- In choosing an element type, first-order solid continuum hexahedral elements with reduced-integration and hourglass controls are recommended. They are robust at handling the many sources of nonlinearity that arise in DSIF. However, for improved accuracy, it is important to create a structured mesh that limits element distortion at finite strains. Also, more studies using different finite element technologies should be explored, such as B-bar formulations [16] and hexahedral thick-shell elements.
- The ratio between the global internal energy and global kinetic energy of the sheet should not be used to determine whether a simulation of DSIF is quasi-static. This energy ratio is not applicable because the nature of DSIF involves highly localized deformations. Moreover, the average contact forces are also not a good indicator of whether a DSIF simulation is quasi-static. In future work, it would be worthwhile to formulate a local energy ratio by defining a characteristic length, which could be based on the area of the contact patch between the tools and the sheet. For now, the surest method of evaluating the effects of artificial acceleration is to directly examine the changes in the mechanics, such as the stress triaxiality and Lode angle parameter.

6. References

1. Moser N, Zhang Z, Ren H, Zhang H, Shi Y, Ndip-Agbor EE, Lu B, Chen J, Ehmann KF, Cao J (2016) Effective forming strategy for double-sided incremental forming considering in-plane curvature and tool direction. CIRP Ann - Manuf Technol 65:265–268. https://doi.org/10.1016/j.cirp.2016.04.131
2. Lu B, Fang Y, Xu DK, Chen J, Ai S, Long H, Ou H, Cao J (2015) Investigation of material deformation mechanism in double side incremental sheet forming. Int J Mach Tools Manuf 93:37–48. https://doi.org/10.1016/j.ijmachtools.2015.03.007
3. Li Y, Chen X, Liu Z, Sun J, Li F, Li J, Zhao G (2017) A review on the recent development of incremental sheet-forming process. Int J Adv Manuf Technol 92:2439–2462. https://doi.org/10.1007/s00170-017-0251-z
4. Moser N, Pritchet D, Ren H, Ehmann KF, Cao J (2016) An efficient and general finite element model for double-sided incremental forming. J Manuf Sci Eng 138: https://doi.org/10.1115/1.4033483
5. Smith J, Malhotra R, Liu WK, Cao J (2013) Deformation mechanics in single-point and accumulative double-sided incremental forming. Int J Adv Manuf Technol 69:1185–1201. https://doi.org/10.1007/s00170-013-5053-3

6. Esmaeilpour R, Kim H, Park T, Pourboghrat F, Mohammed B (2017) Comparison of 3D yield functions for finite element simulation of single point incremental forming (SPIF) of aluminum 7075. Int J Mech Sci 133:544–554. https://doi.org/10.1016/j.ijmecsci.2017.09.019
7. Henrard C, Bouffioux C, Eyckens P, Sol H, Duflou JR, Van Houtte P, Van Bael A, Duchêne L, Habraken AM (2011) Forming forces in single point incremental forming: Prediction by finite element simulations, validation and sensitivity. Comput Mech 47:573–590. https://doi.org/10.1007/s00466-010-0563-4
8. Ren H, Moser N, Zhang Z, Ndip-Agbor E, Smith J, Ehmann KF, Cao J (2015) Effects of tool positions in accumulated double-sided incremental forming on part geometry. J Manuf Sci Eng Trans ASME 137:1–8. https://doi.org/10.1115/1.4030528
9. Belytschko T, Liu WK, Moran B, Elkhodary K (2014) Nonlinear Finite Elements for Continua and Structures, 2nd ed. Wiley
10. Wriggers P (2008) Nonlinear Finite Element Methods. Springer-Verlag Berlin Heidelberg
11. Malhotra R, Xue L, Belytschko T, Cao J (2012) Mechanics of fracture in single point incremental forming. J Mater Process Technol 212:1573–1590. https://doi.org/10.1016/j.jmatprotec.2012.02.021
12. Shin J, Bansal A, Nath M, Cheng R, Banu M, Taub A, Martinek B (2018) Prediction of Negative Bulge in Two Point Incremental Forming of an Asymmetric Shape Part. J Phys Conf Ser 1063:012057. https://doi.org/10.1088/1742-6596/1063/1/012057
13. Flanagan DP, Belytschko T (1981) A Uniform Strain Hexahedron and Quadrilateral with Orthogonal Hourglass Control. Int J Numer Methods Eng 17:679–706. https://doi.org/10.1002/nme.1620170504
14. Lee PS, Bathe KJ (2005) Insight into finite element shell discretizations by use of the “basic shell mathematical model.” Comput Struct 83:69–90. https://doi.org/10.1016/j.compstruc.2004.07.005
15. Tekkaya AE (2000) State-of-the-art of simulation of sheet metal forming. J Mater Process Technol 103:14–22. https://doi.org/10.1016/S0924-0136(00)00413-1
16. Hughes T (1980) Generalization of Selective Integration Procedures to Anisotropic and Nonlinear Media. Int J Numer Methods Eng 15:1413–1418. https://doi.org/10.1002/nme.1620150914

A Temperature Instability in 4 K Cryocooler Regenerators Caused by Real Fluid Properties

R. Snodgrass[1], V. Kotsubo[2], J. Ullom[1,2], and S. Backhaus[1,2]

[1]National Institute of Standards and Technology, Boulder, CO 80305
[2]Department of Physics, University of Colorado Boulder, Boulder, CO 80309

ABSTRACT

We report temperature profile measurements from a densely instrumented, commercial pulse tube refrigerator. Azimuthal temperature differences of 15 K were measured across its 3 cm diameter regenerator which was operated at cold end temperatures below 10 K. These asymmetries may appear and disappear with just 0.1 K changes to the cold end temperature, suggesting a potential thermofluid instability. Experiments and analysis suggest that real fluid properties of helium at low temperatures may be the driving mechanism of the instability. We sketch the beginnings of a linear perturbation analysis and show that small changes to regenerator temperature profiles are reinforced by accompanying changes to the component of power flow due to real fluid properties, particularly at temperatures less than 9 K. Our measurements show that temperature asymmetries are specific to particular sections within the regenerator and negatively affect cooling power at the cold end.

INTRODUCTION

In low-temperature cryocoolers, real fluid properties[1,2] and finite solid heat capacity[3,4] can cause the temperature profile of regenerators to be far from linear.[5] For example, the temperature profile is usually flat at the cold end of the regenerator because near 4 K most power is carried by terms that are independent of temperature gradient. While the effect of real fluid properties on temperature profile has been considered previously in the literature, the possibility of instabilities arising from these properties has not yet – to the best of our knowledge – been observed or studied.

Instabilities exist in a variety of thermoacoustic machines. In refrigerators with large diameter regenerators, transverse, spatial variations in mass streaming and mean temperature couple[6] and reinforce each other, leading to an increased thermal load on the cold heat exchanger and reduced efficiency. Fluid diodes used in feedback pulse tube refrigerators[7] may be unable to suppress streaming in the direction of the acoustic axis if operated outside of a region of stability. Inverted pulse tubes have been compared to classical pendulums,[8,9] where gravity-driven convection may be suppressed if the acoustics are driven at high enough frequency. Each of these instabilities has been studied in the context of ideal gas working fluids. Below approximately 30 K, helium is strongly non-ideal, and its real fluid properties may also lead to instabilities.

Snodgrass, Ryan; Kotsubo, Vincent; Ullom, Joel; Backhaus, Scott. "A Temperature Instability in 4 K Cryocooler Regenerators Caused by Real Fluid Properties." Presented at 21st International Cryocooler Conference. December 07, 2020 - December 10, 2020.

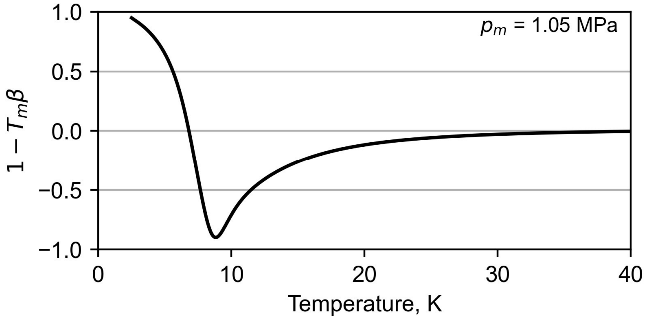


Figure 1. One minus mean temperature (T_m) multiplied by the thermal expansion coefficient (β) of helium-4 at a mean pressure of 1.05 MPa.

An incomplete but intuitive understanding of an instability driven by real fluid properties is gained by considering the total power flow through the regenerator. One component of the total power is $\dot{H}_\beta = (1 - T_m\beta)\dot{E}_2$. $\dot{E}_2$ is the acoustic power, which is always directed from the warm end of the regenerator towards the cold end; T_m is mean temperature; and β is the thermal expansion coefficient. In the ideal gas regime (roughly above 30 K) $1 - T_m\beta = 0$, and $\dot{H}_\beta$ does not carry significant power. In the real fluid regime $1 - T_m\beta$ is different from 0, and this term becomes important. Figure 1 displays a plot of $1 - T_m\beta$ versus T_m at 1.05 MPa. At this pressure, this mode of power flow is positive for $T_m < 6.8$ K and appears as a heat load on the cold heat exchanger, reducing the available cooling power. It is negative for $T_m > 6.8$ K and enhances the transport of heat away from the cold region toward the warm region. These strong shifts in the ability to move energy either toward or away from the cold end as temperature changes – from this and other related power flow terms – are a primary source of thermofluid instability.

As an example, consider an axial section of the regenerator at 15 K during cooldown to 4 K. If the temperature of one azimuthal half of the regenerator were increased in temperature to 16 K and the other azimuthal half were decreased to 14 K, the colder half will cooldown at a quickening pace because $(1 - T_m\beta)\dot{E}_2$ is now more negative (increasing cooling power). In contrast, the half at 16 K cools at a relatively slower rate. These dynamic changes to the power flow will lead to diverging temperatures in the two halves of the regenerator that will continue until other power flow effects lead to saturation and a steady state.

In the remainder of this paper, we detail our experimental systems and procedures, describe our observations of a potential instability leading to azimuthal temperature asymmetries in the second-stage regenerator of a pulse tube refrigerator, and provide a preliminary perturbation analysis explaining how real fluid effects may be the mechanism driving the instability.

EXPERIMENTAL SYSTEM AND PROCEDURES

Figure 2 displays a schematic of our experimental system. Additional details of our experiment can be found in our group's second paper[10] for this conference. The system consists of a two-stage, commercial pulse tube refrigerator operating near 1.4 Hz. Our focus is on the second-stage regenerator that nominally spans temperatures from 40 K to 60 K at its warm end to 2.5 K to 10 K at the cold end. The diameter of the second-stage regenerator is close to 3 cm. We have densely instrumented the second-stage with 18 silicon diodes to measure the temperature at 9 axial locations along the regenerator using 2 diodes per axial location located on opposite sides of the regenerator. Each diode in a pair is mounted on one of two opposing copper clamps that mate around the regenerator shell. The two halves of each clamp are drawn together with two 4-40 stainless bolts but do not directly touch, so they are mostly thermally isolated from each other. The contact area between each clamp and the regenerator shell is an azimuthal band only 1.27 mm tall and slightly less than 180 degrees around the circumference. This small thermal contact patch minimally affects the temperature gradient along the regenerator. Two additional diodes are bolted

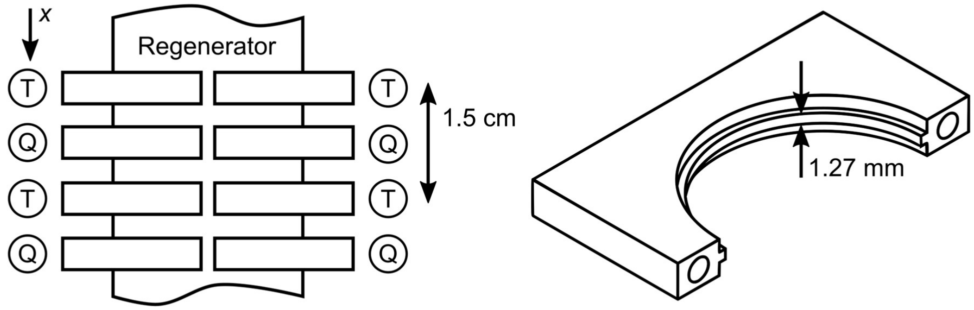


Figure 2. Schematic of the experimental test setup. Two opposing pieces of copper clamp around the second-stage regenerator tube (side view, left, and isometric view of one piece, right). A diode thermometer (T) is placed on each opposing clamp half so that the temperature of different regenerator halves can be compared. (Q) shows where we have the capability to apply intermediate heat, although we did not use this feature for the experiments presented in this paper.

directly to the copper of the warm and cold heat exchangers. All diode thermometers were calibrated to an accuracy of ± 25 mK at temperatures < 25 K and to ± 75 mK for higher temperatures. At the warm and cold heat exchangers we have placed heaters, which are integrated into a feedback system to control the warm (T_w) and cold end (T_c) temperatures to better than ± 20 mK. These control loops are utilized in all experiments reported here.

In a typical experiment, the cryocooler is cooled down, the control loops for T_w and T_c are engaged, and the system is allowed to come into steady state. We then change T_c in 0.1 K increments in a step-like fashion, measuring temperature at all locations at a rate of roughly 0.5 Hz. We wait to proceed with the next temperature change until steady state is reached: this can take about 10 minutes if the temperature profile does not change drastically, or on the order of one hour if transitioning into or out of large temperature asymmetries. We scan T_c in 0.1 K steps near instability conditions in both directions, i.e. cooling down and warming up.

OBSERVATIONS

Figure 3a displays the maximum temperature asymmetry at any axial location along the regenerator when T_c is swept from 7.6 K to 6.2 K in increments of 0.1 K while holding T_w at 46 K.

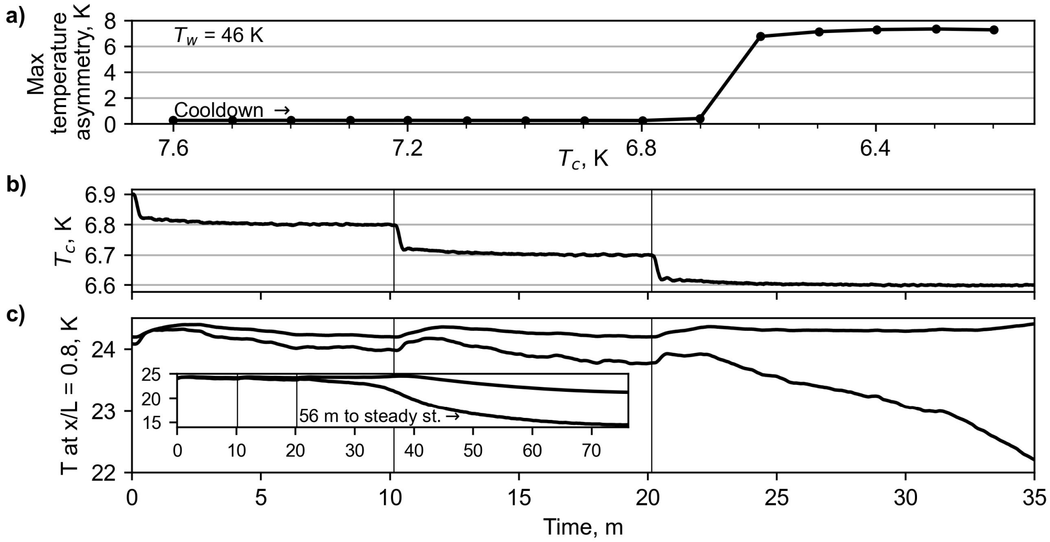


Figure 3. Temperature asymmetry forming during cooldown. a) Maximum steady state temperature difference between opposing thermometers at any axial location. Transient data for a subset of the data from a) are shown in b) and c). The cold end temperature is shown in b), and the temperature of opposing thermometers is shown for a particular regenerator location in c). Vertical lines show when T_c was changed.

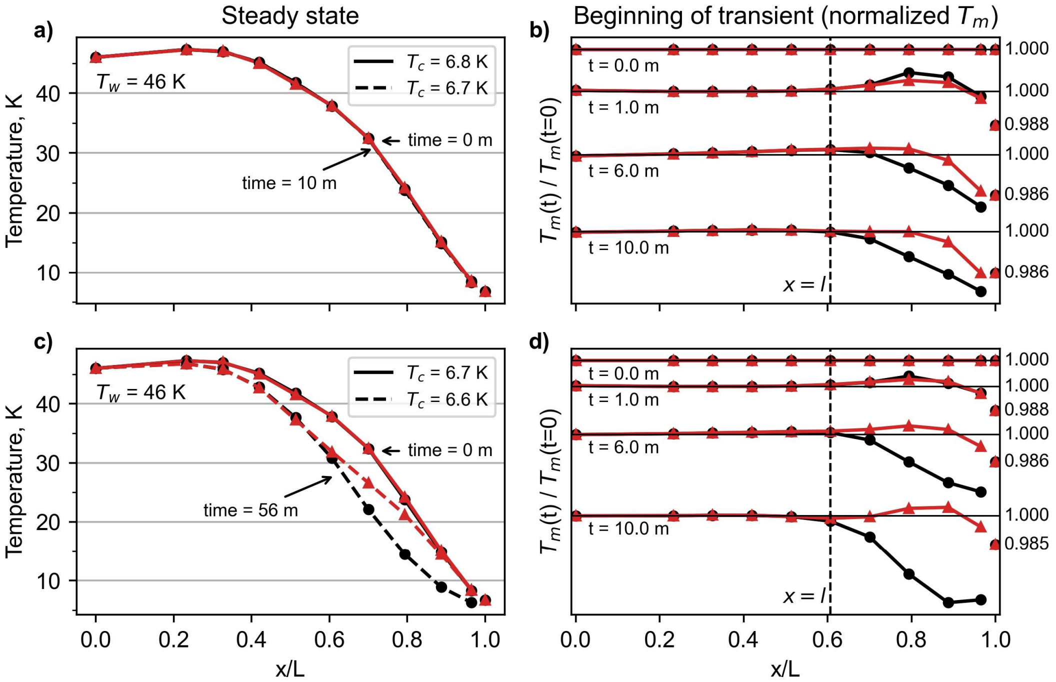


Figure 4. Cooling down the cold end in 0.1 K increments, demonstrating stability (top row) and instability (bottom row). Black lines with circular markers give the temperature of one azimuthal half of the regenerator; red lines with triangular markers give the opposite half. a) and c) give the steady state profiles, while b) and d) show the normalized temperature profile over time at four different timepoints. Straight horizontal lines on the right subplots give $T_m(t)/T_m(t=0) = 1$. The normalized distance along the regenerator is x/L, and $x = l$ is identified near Eq. 1. The profile shown in c) takes about 56 minutes to reach steady state after T_c change. The large temperature changes at 56 minutes are not shown in d) since they would obscure the important dynamics in the first 10 minutes after the change in T_c.

This maximum temperature difference is our measure of the asymmetry between the two halves of the regenerator. At higher values of T_c, the two halves of the regenerator display negligible asymmetry with small differences between the two halves of approximately 0.25 K. This asymmetry may be attributable to manufacturing variability. The dynamics of the cold end temperature following a 0.1 K step-like decrease in the control loop set point for T_c are shown in Figure 3b. Figure 3c shows the dynamic temperature evolution of the pair of thermometers on the clamps at $x/L = 0.8$, where x is the distance along the regenerator axis starting at the warm heat exchanger and L is the total length of the regenerator. While in the stable regime with no large temperature asymmetry, the temperature profile in the middle of the regenerator reaches steady state within about 10 minutes after incrementing T_c (first two T_c steps in Figure 3b). When the cold end drops further to 6.6 K (third T_c step in Figure 3b), the temperature difference between the pair of thermometers at $x/L = 0.8$ grows from much less than 1 K to about 7 K over the course of about an hour. The sudden onset of this asymmetry is an indication of a thermofluid instability in the second-stage regenerator.

Figure 4 shows two T_c increments from Figure 3 but over the entire temperature profile, for $x/L = 0$ to $x/L = 1$. The steady state profiles for the last T_c step before instability are shown in Figure 4a (note: no noticeable change in temperature profile), and the dynamic transition between these profiles is shown in Figure 4b. Figure 4c and 4d show the same data for the first unstable step in T_c (note: a very large change in temperature profile). For the dynamic data in Figures 4b and 4d, the time-dependent temperatures are normalized by the steady-state temperature before the change in T_c to enhance the visibility of the dynamics. One minute after T_c is stepped down, the normalized profile looks similar for the stable step to 6.7 K (Figure 4b) as for the unstable step to 6.6 K (Figure 4d). In Figure 4d at six minutes after T_c is changed, the initial shape of the

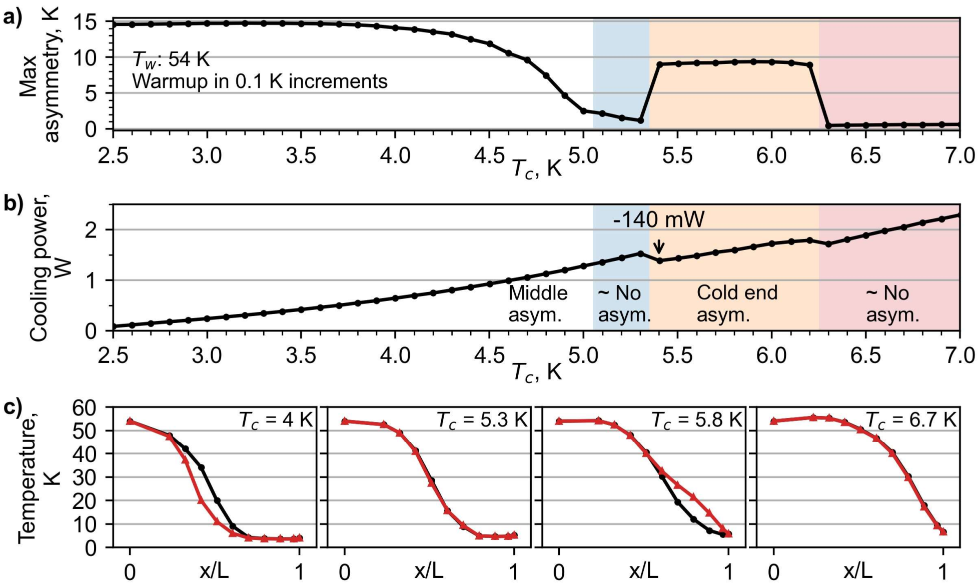


Figure 5. a) Maximum steady state temperature asymmetry as T_c is incremented up in 0.1 K steps while T_w is regulated to 54 K. b) Cooling power at the cold end for the same experiments in a). Colored shading shows the characteristic shape of the asymmetry for each T_c. c) Example temperature profiles for each shaded region in a) and b). Red and black lines represent the different halves of the regenerator.

temperature profile that eventually leads to the larger asymmetry is clearly seen. In the next section, we will use an approximation to this observed shape as the start of a perturbation analysis.

Figure 5a shows maximum temperature asymmetry between the two regenerator halves over a wider range of T_c than presented in Fig. 3. For the data in Fig. 5, the warm end is regulated to 54 K (somewhat warmer than in Figs. 3 and 4) and the T_c increments are +0.1 K (instead of –0.1 K as in Figs. 3 and 4). Over this wider range of T_c, there are multiple regions of large asymmetry with bordering regions of negligible asymmetry. Figure 5c shows the steady-state temperature profiles at values of T_c in each of these regions, which demonstrate that the temperature asymmetry can exist within different parts of the regenerator. At the coldest temperatures, the asymmetry exists in the middle third of the regenerator. The asymmetry gradually fades as T_c approaches about 5 K. At T_c = 5.3 K, the asymmetry almost vanishes with the maximum temperature difference between halves at just 1.1 K (versus 14.7 K at T_c = 3.2 K). Raising T_c by just 0.1 K brings about a new asymmetry of about 9 K, but now the asymmetry resides in the coldest third of the regenerator. At T_c = 6.3 K and above the asymmetry at the cold end disappears just as abruptly as it appeared.

The cooling power $\dot{Q}_c$ at the cold heat exchanger for each T_c is shown in Fig. 5b. There are discontinuities in $\dot{Q}_c$ when the asymmetry at the cold end appears and disappears. When raising T_c from 5.3 K to 5.4 K, we see a decrease in cooling power of 0.14 W, although we would normally expect cooling power to rise with increasing T_c. We do not observe any changes to cooling power at the cold end due to changes in the asymmetry in the middle of the regenerator, although it is possible that this asymmetry could affect intermediate cooling (cooling available along the regenerator itself – discussed more in our companion paper[10] for this conference).

The copper clamps we have mounted to the regenerator for temperature measurement likely cause the asymmetry to take a certain azimuthal orientation. Recall Fig. 2, showing opposing copper clamps, each of which contact the regenerator around a circumferential band slightly less than 180 degrees. Since the copper clamps at different x locations are oriented identically (i.e. they are not randomly rotated), we expect that the clamps force the asymmetry to take the same orientation. We have made measurements where only two pairs of copper clamps were mated to the regenerator, and these pairs were modified so that each piece of copper contacted the regenerator over a small number of degrees. With the two pairs of small contact clamps, the asymmetry appeared at similar temperatures (differences in onset temperature between 0.1 K and 1 K versus experiments with many clamps) and produced temperature differences of similar magnitude to the measurements with many clamps. In the two pair configuration, we did observe signs that the asymmetry could rotate azimuthally (as T_c was incremented).

PRELIMINARY INSTABILITY ANALYSIS

The objective of our preliminary analysis is to reveal possible destabilizing influences that lead to temperature asymmetry. Our focus here will only be on the contribution from real fluid properties: we reserve a full analysis for future work. We start by considering a steady state symmetric temperature profile, i.e., the same $T_m(x)$ in both regenerator halves. At time $t = 0$, we introduce a small temperature perturbation into the left half $\delta T_{m,left}(x, t = 0)$ and right half $\delta T_{m,right}(x, t = 0)$ of the regenerator (we are not concerned with the source of the perturbation). This perturbation causes changes to the power flows through the regenerator, which themselves cause changes to $\delta T_{m,left}(x, t)$ and $\delta T_{m,right}(x, t)$. The temperature and power flow perturbations become dynamic, and if the power flow perturbation reinforces the temperature perturbation, then the perturbation is unstable and grows into an asymmetry.

Analytical computation of the self-consistent shape of the temperature perturbation is beyond the scope of this work. Instead, we leverage the observed shape of the transient temperatures following a step change in T_c (see Fig. 4d at t = 6 minutes). In this preliminary analysis, we estimate the shape as

$$\begin{aligned} &\delta T_{m,left}(x,t) = -\delta T_{m,right}(x,t) = \delta T_0(t)\sin\left(\frac{\pi}{2}\frac{x-l}{L-l}\right) && \text{for } \; l < x < L \\ &\delta T_m(x,t) = 0 && \text{for } \; 0 < x < l, \end{aligned} \tag{1}$$

where x is the distance along the regenerator axis, and $\delta T_0(t)$ is the amplitude of the perturbation, which may grow or decay with time depending on the relative balance of stabilizing and destabilizing effects. The perturbation begins at $x = l$ where the temperatures are the same, but where there is a difference in the temperature gradient between the left and right halves. For $x > l$ there is a temperature difference between the left and right halves that grows and takes on its largest value at the end of the regenerator at $x = L$, where the temperature gradients are the same.

The data in Fig. 4d show similar behavior with the temperature deviation between left and right beginning near $x/L \sim 0.6$ (i.e., at $x = l$) and reaching a maximum at x/L = 0.91 or x/L = 0.97. The last data point in Fig. 4d at x/L = 1 is recorded by a single diode that is bolted to the cold heat exchanger itself. Significant mixing of the helium as it flows through the cold heat exchanger likely homogenizes the temperature at that point, which is why in Fig. 4 we have shown breaks in the temperature profiles at the cold end.

Our analysis of the power flows will use the thermoacoustic framework pioneered by Rott[11] and further developed by Swift[12]. Consider a simplified version of the total power equation:

$$\dot{H}_2(x) = \frac{1}{2}\mathrm{Re}\left[p_1\widetilde{U_1}\left(1 - \frac{T_m\beta\left(f_\kappa - \widetilde{f_\nu}\right)}{(1+\epsilon_s)(1+\sigma)\left(1-\widetilde{f_\nu}\right)}\right)\right] + \dot{H}_T + \dot{H}_m + \dot{H}_\kappa, \tag{2}$$

where p_1 and U_1 are complex amplitudes of the oscillating pressure and volumetric flowrate, tilde represents the complex conjugate, σ is the Prandtl number, and $\epsilon_s = \phi\rho c_p/(1-\phi)\rho_s c_s$ is the ratio of the fluid to solid volumetric heat capacity in the regenerator ($\phi, \rho, c_p, \rho_s, c_s$ are the porosity, fluid density, fluid isobaric specific heat, solid density, and solid specific heat, respectively). The terms in $\dot{H}_2$ with subscripts T, m, and κ are power flows carried by oscillating flow along a temperature gradient, steady flow (streaming), and thermal conduction, respectively. We do not write out those terms explicitly because here we will restrict our analysis to the first term in Eq. 2.

The thermoacoustic functions f_κ and f_ν in Eq. 2 depend on the ratio of the hydraulic radius to the thermal penetration depth, i.e., r_h/δ_κ.[12] This ratio is expected to be of order 0.1 in highly effective low-temperature regenerators. Using expressions for f_κ and f_ν in a parallel plate geometry as an approximation to porous media, we expand those expressions for small r_h/δ_κ and simplify the first power flow term in Eq. 2 to

$$\dot{H}_2(x) \approx (1 - T_m\beta)\dot{E}_2 + \left(\frac{\epsilon_s}{1+\epsilon_s}\right) T_m\beta\dot{E}_2 + \dot{H}_T + \dot{H}_m + \dot{H}_\kappa. \tag{3}$$

This form of the power flow equation splits the first term in Eq. 2 into components that do and do not depend on ϵ_s. In the temperature regime of interest ($T_m < 30$ K), $1 - T_m\beta$ is significantly different than 0. In practical cryocoolers, ϵ_s can be of order 1 across some or all of the second-stage regenerator. Therefore, we expect the two terms proportional to acoustic power $\dot{E}_2$ in Eq. 3 to carry a significant fraction of the total power flow in the region of the potential instability in Fig. 4d. The effect of the temperature perturbation in Eq. 1 on these power flow terms plays a large role in determining if an instability arises.

Now we will continue our preliminary analysis by focusing our attention to the term that carries power when real fluid properties are most important, i.e. $\dot{H}_\beta = (1 - T_m\beta)\dot{E}_2$. The response of this power flow term to a temperature perturbation is

$$\delta\dot{H}_\beta = \delta T_m\, d\dot{H}_\beta/dT_m, \tag{4}$$

where the total derivative must be used because $T_m, \beta,$ and $\dot{E}_2$ are all dependent on T_m. The magnitude and time phasing of the oscillating (acoustic) pressure throughout the regenerator are typically constant, so that p_1 can be taken as real, and the acoustic power can be approximated as $\dot{E}_2 = p_1\mathrm{Re}[U_1]/2$, so that the total derivative is

$$\frac{d\dot{H}_\beta}{dT_m} = \frac{\partial\dot{H}_\beta}{\partial T_m} + \frac{\partial\dot{H}_\beta}{\partial\mathrm{Re}[U_1]}\frac{\partial\mathrm{Re}[U_1]}{\partial T_m}. \tag{5}$$

Conservation of mass requires $\rho\mathrm{Re}[U_1] =$ constant. Differentiating this expression and using the definition $-\beta\rho = \partial\rho/\partial T_m|_p$, Eq. 5 is expressed

$$\frac{d\dot{H}_\beta}{dT_m} = \dot{E}_2\left[\frac{-\partial(T_m\beta)}{\partial T_m} + (1 - T_m\beta)\beta\right]. \tag{6}$$

Combining this expression with Eq. 4, we can compute the power flow perturbation created by the temperature perturbation.

Next, we compute the joint time evolution of the temperature and power flow perturbations. Assuming that the fluid and solid are at the same temperature T_m, the perturbed energy equation requires

$$[\phi\rho c_p + (1-\phi)\rho_s c_s]A\frac{\partial\delta T_m}{\partial t} = -\frac{\partial\delta\dot{H}_\beta}{\partial x}, \tag{7}$$

where A is the cross-sectional area of the regenerator. Equation 7 reveals that these perturbations are not in steady state. Variations of the power flow perturbation along the regenerator axis will result in time varying accumulations of the energy in different regions of the regenerator, which drives time varying temperatures in those regions. A complete solution to Eq. 7 is beyond the scope of this preliminary analysis. Instead, we are guided by a simplified analysis used in other stability studies[7]. We begin by substituting $\delta T_m(x,t)$ from Eq. 1 into Eq. 7 and integrating from $x = l$ to $x = L$, which yields

$$C\frac{\partial\delta T_0(t)}{\partial t} = \delta\dot{H}_\beta(l) - \delta\dot{H}_\beta(L) \equiv \Delta\delta\dot{H}_\beta = \dot{E}_2\left[\frac{\partial(T_m\beta)}{\partial T_m} - (1 - T_m\beta)\beta\right]_{x=L}\delta T_0(t). \tag{8}$$

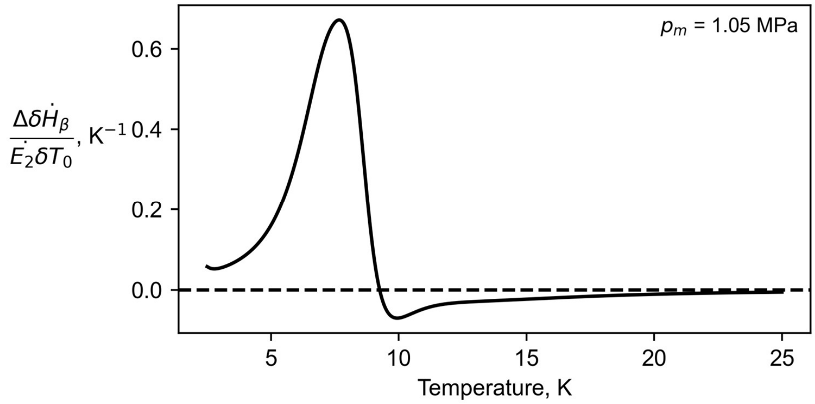


Figure 6. The power perturbation due to real fluid properties in the section of the regenerator between $x = l$ and $x = L$, normalized by acoustic power and the amplitude of the temperature perturbation.

The constant C in Eq. 8 is an average of the combined fluid and solid heat capacity weighted by the spatial dependence of the temperature perturbation. Equation 8 states the physically intuitive result that the amplitude of the temperature perturbation between $x = l$ and $x = L$ grows in time if the imbalance in the power flows at the two ends of the region create a net of power into the region, i.e., if $\delta\dot{H}_\beta(l) - \delta\dot{H}_\beta(L) > 0$. The difference $\delta\dot{H}_\beta(l) - \delta\dot{H}_\beta(L)$ in Eq. 8 is rewritten by setting $\delta\dot{H}_\beta(l) = 0$ (because $\delta T_m(x,t) = 0$ at $x = l$) and computing $\delta\dot{H}_\beta(L)$ by evaluating Eq. 6 at $x = L$, where $\delta T_m(x,t) = \delta T_0(t)$.

From Eq. 8, the amplitude of the perturbation $\delta T_0(t)$ grows exponentially in time when $\Delta\delta\dot{H}_\beta > 0$, where $\Delta\delta\dot{H}_\beta$ is given in the last term on the right-hand side of Eq. 8. Figure 6 displays $\Delta\delta\dot{H}_\beta$ as a function of the cold end temperature, i.e. $T_c = T_m(x = L)$. For T_c less than about 9 K, $\Delta\delta\dot{H}_\beta > 0$. At these temperatures, we predict that the original temperature perturbation grows due to perturbations to total power and is unstable. From this analysis, real fluid properties are a destabilizing mechanism at $T_m < 9$ K; however, analysis of the effect of $\delta T_m(x,t)$ on the other power flow terms is needed to make a more accurate prediction of the cold end temperature where perturbations become unstable.

DISCUSSION

Although we have shown how one term in the total power equation may promote temperature instability at low temperatures, each term in the power equation will have its own stabilizing or destabilizing contribution. For example, we have performed preliminary work that suggests the finite heat capacity term in the total power equation (i.e. the term containing ϵ_s in Eq. 3) largely promotes stability over the same temperatures that $\dot{H}_\beta$ promotes instability. A stabilizing influence from this term will push the threshold for transition to instability to a temperature colder than 9 K, which would be more in line with the transition temperature observed in our experiments. Because this ϵ_s term is highly dependent upon the regenerator material, it is likely that each cryocooler will exhibit this low-temperature instability differently.

Stabilizing effects from finite solid heat capacity may also be responsible for the compartmentalization of the asymmetry we see in Fig. 5c. Since it is common for low-temperature regenerators to be constructed from multiple porous media, it's quite possible that we do not observe asymmetries between the middle third and coldest third of the regenerator because this location is a transition between materials. The lack of asymmetry might be explained by differences in the stabilizing effects from materials of dissimilar heat capacity, or by high-thermal-conductivity materials separating regenerator materials and promoting azimuthal temperature uniformity.

It is also important to note that the asymmetry onset is dependent upon more than just T_c. Consider the experiment in Fig. 5, where the cold end asymmetry is not present at $T_c \geq 6.3$ K, while Fig. 3 shows that the same asymmetry is not present at $T_c \geq 6.7$ K. The warm end temperatures for these experiments were 54 K and 46 K, respectively. This difference in onset temperature likely results from a difference in the stabilizing or destabilizing contributions from dT_m/dx terms in the total power equation ($\dot{H}_T$ and $\dot{H}_\kappa$ in Eq. 3). We have performed experiments with T_w up to 62 K. We generally observe stability at lower T_c with increasing T_w, suggesting that at least one of the dT_m/dx terms in Eq. 3 are stabilizing.

CONCLUSION

We have observed large temperature asymmetries (azimuthal temperature differences) in a low-frequency, low-temperature pulse tube refrigerator. The results presented here show that the asymmetry can be 15 K across a second-stage regenerator of just 3 cm diameter, which is a significant portion of the second-stage temperature span (~ 45 K). In other experiments not presented here, we have observed the asymmetry approach 19 K. The asymmetry appears to be restricted to certain regions of the regenerator – either the third of the regenerator closest to the cold end, or the middle third – which we believe to be a result of the transitions between different porous media. The onset of the asymmetry at the cold end depends sensitively on temperature (i.e. with changes to end conditions of just 0.1 K), while the asymmetry in the middle grows or disappears more gradually with changes to end conditions.

We also have presented the beginning of a linear perturbation analysis using transient measurements of the temperature profile to estimate the shape of the temperature perturbation that may lead to instability. We showed that the power flow term that depends strongly on the real fluid properties of helium ($\dot{H}_\beta$) promotes temperature instability at temperatures less than about 9 K. Some or all of the terms of the total power equation beside $\dot{H}_\beta$ likely promote stability; a complete perturbation analysis must consider all terms to fully explain stability criteria. Besides studying these other power terms and their influence on stability, we are currently in the process of performing a sweep of T_w and T_c to map out the stable/unstable state space.

Our work also demonstrates that accurately measuring the temperature profile of low-temperature cryocooler regenerators is not possible without considering azimuthal temperature variation.

ACKNOWLEDGMENTS

R. Snodgrass acknowledges support from a National Research Council Postdoctoral Fellowship. We would like to thank Greg Swift for stimulating discussion and leadership, and for assistance with the thermoacoustic software DeltaEC.

Figures were generated using Matplotlib and material properties were gathered using CoolProp[13], an open-source thermophysical library available as a Python wrapper.

Contribution of NIST, not subject to copyright.

REFERENCES

1. de Waele, A.T.A.M., Xu, M.Y., and Ju, Y.L., "Nonideal-gas effect in regenerators," *Cryogenics*, vol. 39, no. 10, (1999), pp. 847–851.
2. Cao, Q., Qiu, L., and Gan, Z., "Real gas effects on the temperature profile of regenerators," *Cryogenics*, vol. 61, (2014), pp. 31–37.
3. de Waele, A.T.A.M., "Finite heat-capacity effects in regenerators," *Cryogenics*, vol. 52, no. 1, (2012), pp. 1–7.
4. Wang, C., "Numerical analysis of 4 K pulse tube coolers: Part I. Numerical simulation," *Cryogenics*, vol. 37, no. 4, (1997), pp. 207–213.
5. Lang, A., Häfner, H.-U., and Heiden, C., "Systematic Investigations of Regenerators for 4.2K-Refrigerators" in *Advances in Cryogenic Engineering*, Boston, MA: Springer US, pp. 1573–1580, (1998).
6. So, J.H., Swift, G.W., and Backhaus, S., "An internal streaming instability in regenerators," *J. Acoust. Soc. Am.*, vol. 120, no. 4, (2006), pp. 1898–1909.
7. Backhaus, S., and Swift, G.W., "An acoustic streaming instability in thermoacoustic devices utilizing jet pumps," *J. Acoust. Soc. Am.*, vol. 113, no. 3, (2003), pp. 1317–1324.

8. Swift, G.W., and Backhaus, S., "The pulse tube and the pendulum," *J. Acoust. Soc. Am.*, vol. 126, no. 5, (2009), pp. 2273–2284.

9. Carbo, R.M., Smith, R.W.M., and Poese, M.E., "A computational model for the dynamic stabilization of Rayleigh-Bénard convection in a cubic cavity," *J. Acoust. Soc. Am.*, vol. 135, no. 2, (2014), pp. 654–668.

10. Snodgrass, R., Kotsubo, V., Ullom, J., and Backhaus, S., "Leveraging Real Fluid Effects as a Tool for Power Flow Measurements in 4 K Cryocooler Regenerators," *Cryocoolers 21*, (2021).

11. Rott, N., "Thermoacoustics" in *Advances in Applied Mechanics* vol. 20, Elsevier, pp. 135–175, (1980).

12. Swift, G.W., *Thermoacoustics: A Unifying Perspective for Some Engines and Refrigerators*, 2nd ed. Springer International Publishing, (2017).

13. Bell, I.H., Wronski, J., Quoilin, S., and Lemort, V., "Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp," *Ind. Eng. Chem. Res.*, vol. 53, no. 6, (2014), pp. 2498–2508.

Mechanisms of Anelastic Loss in Langasite at Temperatures from 113 K to 1324 K

Ward L. Johnson[1], Yuriy Suhak[2], and Holger Fritze[2]
[1]Applied Chemicals and Materials Division, National Institute of Standards and Technology, 325 Broadway St., MS 647, Boulder, CO 80305, USA
[2]Clausthal University of Technology, Am Stollen 19B, Goslar 38640, Germany
ward.johnson@nist.gov

Summary:

Synthetic piezoelectric crystals with the structure of langasite (LGS) are being pursued for resonant acoustic sensors that can operate at temperatures exceeding the range of conventional piezoelectric materials. The optimization of these crystals is currently focused primarily on minimization of acoustic loss, which degrades signal strength and resolution of sensors. This paper presents analysis and discussion of two sets of measurements of loss of LGS with a combined temperature range of 113 K to 1324 K. Physical mechanisms for the loss include intrinsic phonon-phonon interactions, multiple point-defect relaxations, piezoelectric/carrier loss, contact loss, and, perhaps, dislocation relaxations [1].

Keywords: acoustic loss, piezoelectric sensors, high temperatures, langasite, LGS, langatate, LGT catangasite, CTGS, quartz

Introduction

Traditional piezoelectric sensors are limited to operation at temperatures below several hundred degrees Celsius, because crystal transformations or degradation occur at higher temperatures in common commercially available piezoelectrics [2]. However, substantial research in recent decades has focused on synthesizing and optimizing innovative piezoelectric crystals that can be used in resonant sensors at temperatures exceeding 1000 K [3], including crystals with the structure of langasite ($La_3Ga_5SiO_{14}$, “LGS”), often termed members of the “langasite family.”

The performance of resonators in sensing applications is limited by acoustic loss Q^{-1}. Within the langasite family of piezoelectric crystals, LGS has not been found to have the lowest Q^{-1}. For example, resonators of langatate ($La_3Ga_{5.5}Ta_{0.5}O_{14}$, “LGT”) and langanite ($La_3Ga_{5.5}Nb_{0.5}O_{14}$, “LGN”) at room temperature are reported to have lower loss than similarly manufactured LGS resonators [4]. Crystals of LGT [5] and catangasite ($Ca_3TaGa_3Si_2O_{14}$, “CTGS”) [6] have also been found to have lower Q^{-1} at elevated temperatures (*e.g.*, above 200 °C) than that of any reported LGS specimen.

Despite the less-than-stellar quality factor of LGS, the available data on this material currently provide unique information on physical mechanisms that contribute to loss in crystals in the langasite family. Specifically, the range of temperatures over which Q^{-1} in LGS has been measured in the low megahertz range is exceptionally broad, enabling identification of contributions to the loss ranging from the small intrinsic loss associated with phonon-phonon interactions to conductivity-related loss five orders of magnitude larger at elevated temperatures. Analysis and discussion of these LGS data are the focus of this paper.

Results and Discussion

Figure 1(a) shows measurements of Q^{-1} of two Y-cut LGS crystals grown by different manufacturers [2,5]. Measurements on one crystal in vacuum were acquired from 113 K to 752 K with noncontacting electrodes at the National Institute of Standards and Technology (NIST, U.S.A.). Measurements on the other crystal in air were acquired from 309 K to 1324 K with Pt surface electrodes at Clausthal University of Technology (TUC, Germany).

This figure also shows the maximum Q^{-1} at 10 MHz reported for LGS at room temperature with noncontacting electrodes [4]. This Q^{-1} is an order of magnitude smaller than that measured near room temperature on the LGS specimen at NIST, indicating the presence of greater material loss in the NIST specimen. Q^{-1} of the specimen measured at TUC is an additional order of magnitude greater near room temperature. For the purpose of characterizing non-intrinsic contributions to the loss, the relatively high Q^{-1} of the NIST and TUC specimens is advantageous.

Johnson, Ward L.; Suhak, Yuriy; Fritze, Holger. ”Mechanisms of Anelastic Loss in Langasite at Temperatures from 113 K to 1324 K.” Presented at Sensor and Measurement Science International (SMSI 2021). May 03, 2021 - May 06, 2021.

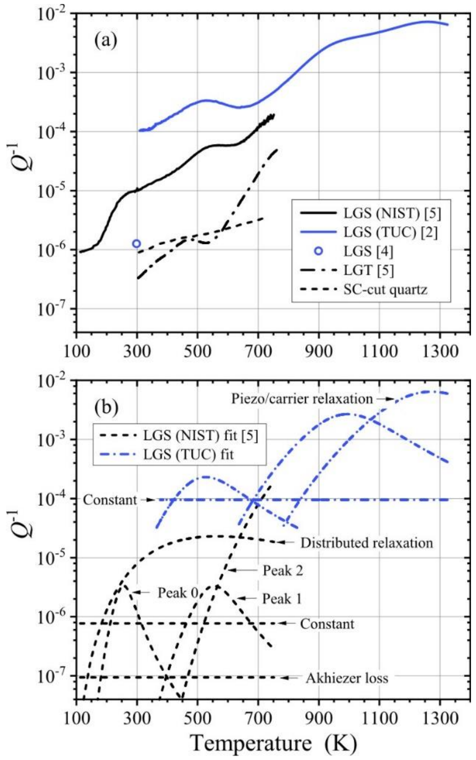


Fig. 1: (a) Q^{-1} of two LGS crystals measured at NIST [5] and TUC [2], an LGS crystal with lowest reported loss [4], an LGT crystal [5], and a swept SC-cut quartz crystal. The resonant frequencies near ambient temperature are, respectively, 6.1 MHz [5], 5.0 MHz [2], 10.0 MHz [4], 6.0 MHz [5], and 10.0 MHz. (b) Contributions to Q^{-1} of the NIST and TUC LGS crystals, determined from least-squares fits.

For comparison, Figure 1(a) also includes data on LGT from 302 K to 759 K [5] and data on swept SC-cut quartz from 306 K to 717 K, all obtained with noncontacting electrodes at NIST.

The LGS data from NIST in Fig. 1(a), along with simultaneously acquired data from two additional harmonics of this crystal, were fit to a function that includes intrinsic phonon-phonon (Akhiezer) loss (approximated as proportional to frequency and independent of temperature) [5], three anelastic point-defect relaxations [7], a constant frequency-independent background, and a broad relaxation consisting of a continuous set of Debye functions [7] with a log-normal distribution of activation energies. The physical mechanism responsible for the last term is hypothesized as arising from dislocations [5], consistent with the fact that no such term is required to fit the data in Fig. 1(a) for LGT, which has much lower dislocation density [5].

Results of fitting of the TUC data at the single measured harmonic are consistent with the NIST results for LGS, with respect to the temperatures of Peaks 1 and 2, considering the difference in resonant frequency. They reveal an additional large peak with a maximum near 1260 K, consistent in form with an expected relaxation involving the motion of charge carriers in acoustically generated piezoelectric fields [5,6]. The fit accurately matches the data without a broadly temperature-dependent term (*e.g.*, distributed relaxation). The constant term, which is inseparable from the Akhiezer term in the absence of measurements of additional harmonics, is two orders of magnitude greater than that determined for the LGS specimen at NIST. This difference is attributed primarily to greater mechanical contact.

Conclusions

Analysis of LGS data from 113 K to 1324 K reveals a number of anelastic loss mechanisms, including intrinsic loss, point-defect relaxations, piezoelectric/carrier relaxation, a constant background, and a broad background that may arise from dislocations. Similar effects have been reported in LGT and CTGS, even when the crystals are state-of-the-art, such as the LGT in Fig. 1(a). Despite the identification of the general nature of loss contributions, the optimization of these innovative piezoelectric materials for applications at elevated temperatures is far from complete. This situation is contrasted, here, with that of swept SC-cut quartz, which shows no evidence for point-defect relaxations over a more limited range of measured temperatures.

[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 bulk acoustic wave sensors, *Meas. Sci. Technol.* 22, 12002 (2011); doi: 10.1088/0957-0233/22/1/012002

[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] Smythe, R. C., Langasite, langanite, and langatate bulk-wave Y-cut resonators, IEEE T. Ultrason. Ferr. 47, 355 (2000); doi: 10.1109/58.827420

[5] 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

[6] 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

[7] Nowick, A. S., and Berry, B. S., *Anelastic Relaxation in Crystalline Solids* (Academic, NY, 1972).

Johnson, Ward L.; Suhak, Yuriy; Fritze, Holger. "Mechanisms of Anelastic Loss in Langasite at Temperatures from 113 K to 1324 K." Presented at Sensor and Measurement Science International (SMSI 2021). May 03, 2021 - May 06, 2021.

Leveraging Real Fluid Effects as a Tool for Power Flow Measurements in 4 K Cryocooler Regenerators

R. Snodgrass[1], V. Kotsubo[2], J. Ullom[1,2], and S. Backhaus[1,2]

[1]National Institute of Standards and Technology, Boulder, CO 80305
[2]Department of Physics, University of Colorado Boulder, Boulder, CO 80309

ABSTRACT

The real fluid properties of helium have a major impact on the thermodynamics of pulse tube and Gifford-McMahon cryocoolers operating below about 30 K. For example, real fluid properties cause the temperature profile in a low-temperature regenerator to be nearly constant at the cold end and allow heat to be applied at warmer, intermediate points along the regenerator axis without affecting cooling power at the cold heat exchanger. We leverage these unique properties and the injection of intermediate heat as a tool for probing and validating the total power equation. As an initial demonstration of this technique, we show how it can be used to measure steady mass flow through the regenerator. We also discuss and demonstrate more advanced measurement protocols that may be used to isolate other terms responsible for power flow in low-temperature regenerators.

INTRODUCTION

The second-stage regenerators of low-frequency pulse tube or Gifford McMahon refrigerators commonly span temperatures between about 50 K and 4 K where thermophysical properties of helium vary dramatically. At these temperatures and at pressures near 1 MPa, helium no longer behaves as an ideal gas but rather as a real fluid. In a real fluid, enthalpy change dh depends upon both temperature and pressure changes:

$$dh = c_p dT + \frac{(1 - T\beta)}{\rho} dp, \tag{1}$$

where T is temperature, c_p is the isobaric specific heat, ρ is density, p is pressure, and β is the thermal expansion coefficient. Mean temperature T_m multiplied by β is highly temperature dependent (Fig. 1). For an ideal gas, $1 - T_m\beta = 0$, but for real fluids $1 - T_m\beta$ can be greater than zero ($T_m < 6.8$ K at mean pressure $p_m = 1.05$ MPa) or less than zero ($T_m > 6.8$ K). Depending on T_m, a pressure change creates a positive or negative change in enthalpy. The variety and unique shape of temperature profiles in low-temperature regenerators[1–3] can be understood by considering $T_m\beta$ and its influence on power flow.

To analyze the power flow we turn to the thermoacoustic framework pioneered by Rott[4] and further developed by Swift.[5] For a real fluid in a regenerator with finite solid heat capacity, the total power equation may be written[6]

Snodgrass, Ryan; Kotsubo, Vincent; Ullom, Joel; Backhaus, Scott. "Leveraging Real Fluid Effects as a Tool for Power Flow Measurements in 4 K Cryocooler Regenerators." Presented at 21st International Cryocooler Conference. December 07, 2020 - December 10, 2020.

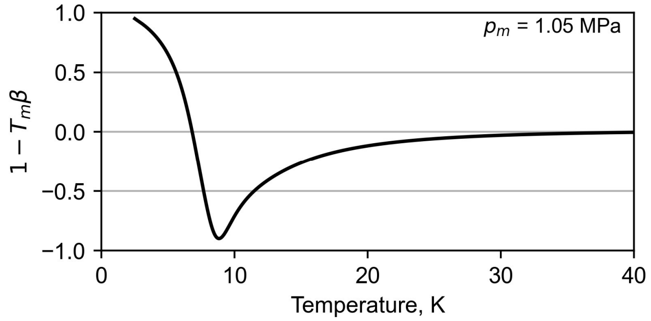


Figure 1. One minus temperature (T_m) multiplied by the thermal expansion coefficient (β) of helium-4 at a mean pressure 1.05 MPa.

$$\begin{aligned}\dot{H}_2(x) = \frac{1}{2}\mathrm{Re}\left[p_1\widetilde{U_1}\left(1-\frac{T_m\beta\big(f_\kappa-\widetilde{f_\nu}\big)}{(1+\epsilon_s)(1+\sigma)\big(1-\widetilde{f_\nu}\big)}\right)\right] + F(x,T_m)|U_1|^2\frac{dT_m}{dx} \\ -\ \big(A_gk+A_sk_s\big)\frac{dT_m}{dx}+\dot{N}h_{mol}.\end{aligned} \tag{2}$$

Here, p_1 and U_1 are complex amplitudes of the oscillating pressure and volumetric flowrate, σ is the Prandtl number, and $\epsilon_s = \phi\rho c_p/(1-\phi)\rho_s c_s$ is the ratio of the fluid to solid volumetric heat capacities in the regenerator (ϕ, ρ_s, c_s are porosity, solid density, and solid specific heat, respectively). A tilde over a parameter signifies the complex conjugate. For brevity in Eq. 2 we have collected parameters that depend on material, geometry, and temperature in the $|U_1|^2 dT_m/dx$ term into $F(x,T_m)$:

$$F(x,T_m) = \frac{\rho c_p}{2A_g\omega(1-\sigma)|1-f_\nu|^2}\mathrm{Im}\left[\widetilde{f_\nu}+\frac{\big(f_\kappa-\widetilde{f_\nu}\big)(1+\epsilon_s f_\nu/f_\kappa)}{(1+\epsilon_s)(1+\sigma)}\right], \tag{3}$$

where f_κ and f_ν are thermoacoustic functions that are known[5] for a variety of regenerator geometries and ω is the angular frequency of the oscillating terms.

In Eq. 2, the first term is connected to the flow of acoustic power $\dot{E}_2 = Re[p_1\widetilde{U_1}]/2$. The $|U_1|^2 dT_m/dx$ term carries power when there is an oscillating volume flow rate along a temperature gradient. The third term accounts for conduction, both through the gas (area A_g and thermal conductivity k) and through the regenerator solid (area A_s and thermal conductivity k_s). The final term is power carried by non-oscillating, steady mass flow (also called streaming) that may be present in geometries that allow it (e.g. a double-inlet pulse tube refrigerator). In this term, $\dot{N}$ is steady molar flow and h_{mol} is the specific molar enthalpy.

To gain more intuition, the $p_1\widetilde{U_1}$ term in Eq. 2 can be simplified by separating it into components that strongly depend only on real fluid effects (i.e. containing $T_m\beta$) and components that depend on finite heat capacity effects (i.e. containing ϵ_s). In realistic regenerators r_h/δ_κ is of order 0.1, and expansion of the thermoacoustic functions f_κ and f_ν in a parallel plate geometry (as an approximation to porous media regenerators), yields

$$\begin{aligned}\dot{H}_2(x) \approx (1-T_m\beta)\dot{E}_2+\left(\frac{\epsilon_s}{1+\epsilon_s}\right)T_m\beta\dot{E}_2+F(x,T_m)|U_1|^2\frac{dT_m}{dx} \\ -\big(A_gk+A_sk_s\big)\frac{dT_m}{dx}+\dot{N}h_{mol}.\end{aligned} \tag{4}$$

Some useful conclusions can already be drawn from Eq. 4. At and near the cold end of a low-temperature regenerator, dT_m/dx is approximately zero. This observation has been made (or predicted) by several other researchers[1–3] and is consistent with our high-resolution measurements reported here. Equation 4 shows that $\dot{H}_2$ at the cold end must be transported by some combination of the two $\dot{E}_2$ terms and by the streaming term. At the cold end of the regenerator, h_{mol} is small,

Snodgrass, Ryan; Kotsubo, Vincent; Ullom, Joel; Backhaus, Scott. "Leveraging Real Fluid Effects as a Tool for Power Flow Measurements in 4 K Cryocooler Regenerators." Presented at 21st International Cryocooler Conference. December 07, 2020 - December 10, 2020.

so the streaming term is generally negligible leaving only the $\dot{E}_2$ terms. In a well-insulated regenerator, the power flow is constant along its axis ($d\dot{H}_2/dx = 0$), and the power flow at the warm end is the same as at the cold end. Near the warm end $(1 - T_m\beta)$ is nearly zero and the regenerator heat capacity is typically large ($\epsilon_s \sim 0$). These properties force most of the power flow to be carried by the other terms in Eq. 4, which results in significant temperature gradients near the warm end. Between the cold and warm ends the temperature profile must adjust to the changing thermophysical properties of the solid(s) and fluid to satisfy $\dot{H}_2(x)$ = constant.

Deep understanding of regenerator performance requires deep understanding of the power equation. Although real fluid effects in these systems have been investigated previously,[1,2] few of these studies have used the thermoacoustic approach, and they have not typically been supported with robust measurements. Our experimental setup – detailed in the next section – was motivated by the desire to validate the thermoacoustic total power equation in realistic low-temperature regenerators and to gain practical knowledge about cryocoolers. Leveraging the properties of real fluids in low-temperature regenerators, we demonstrate techniques to measure steady streaming flow. We also demonstrate more advanced techniques that isolate other power flow terms in Eq. 4.

EXPERIMENTAL METHODS

Our experiments are performed using a two-stage, commercial pulse tube refrigerator operating near 1.4 Hz. Along the second-stage regenerator we attached 20 pairs of copper clamps that serve as mounts for thermometry or as heat injection points (Fig. 2). The two halves of each clamp pair are drawn together with two UNC 4-40 stainless steel screws. The halves do not directly touch, so that they are thermally isolated from each other. The contact area between each clamp half and the regenerator shell is an azimuthal band only 1.27 mm thick and slightly less than 180 degrees around the circumference. This small thermal contact patch minimally effects the temperature profile along the regenerator axis.

Silicon diode thermometers are bolted to half of the clamps along one side of the regenerator providing a spatial resolution of about 1.5 cm (10 diodes total). Two additional diodes are bolted directly to the copper of the warm and cold heat exchangers. Using a single, manufacturer-calibrated diode thermometer, we calibrated all other diode thermometers to an accuracy of ± 25 mK at temperatures less than 25 K and to ± 75 mK for higher temperatures. Of the 20 clamp pairs, the remaining 10 may be used for heat injection. Into each of these clamp-halves we have machined a pocket and have epoxied a resistance heater into the pocket. Unless specified, heat is always applied evenly to both halves to avoid azimuthal variation. We have the capability to apply between 0 and 1.5 W (in 1 mW increments) of heat to each clamp pair. Each pair at different axial locations is independently controlled.

Copper clamps were also attached to the second-stage buffer tube. The clamps were of the same design as the regenerator clamps shown in Fig. 2, only scaled in size to match the buffer tube diameter. Eight silicon diodes measured the temperature of the buffer tube.

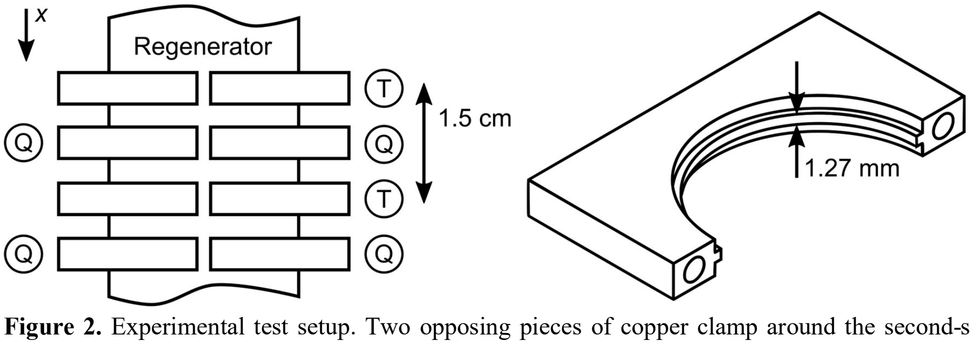


Figure 2. Experimental test setup. Two opposing pieces of copper clamp around the second-stage regenerator tube (side view, left, and isometric view of one piece, right). Half are used to measure temperature (T) and half are used to inject intermediate heat (Q). In most experiments presented here we only measure the temperature on one half of the regenerator.

In a typical experiment the cryocooler is cooled down and heaters and control loops on the warm and cold heat exchangers are used to reach desired warm end and cold end temperatures, T_w and T_c, which are maintained to the target temperatures ± 20 mK. Temperature profiles are recorded when the system reaches steady state. Then, heat is applied to one or more intermediate heat locations and the new temperature profile is recorded when the system reaches steady state again. Intermediate heat $\dot{Q}_{int}$ is used as a tool to change the temperature profile and power flow in sections of the regenerator. Transient temperature data is also recorded for each thermometer at roughly 0.5 Hz. Oscillations in temperature at the frequency of the compressor are filtered by the temperature monitors to give the mean temperature T_m at each location.

RESULTS AND ANALYSIS

Key terms that determine the cooling power at the cold heat exchanger

Even in its simplified form in Eq. 4, the power flow equation is a complex combination of physical effects. We have performed a preliminary set of measurement showing that, under a restricted set of conditions that are often valid, the structure of the power flow equation at the cold heat exchanger is greatly simplified. Later in this paper, the conclusions that we can draw from this simplified form are used to create new techniques to directly probe the power in low-temperature regenerators.

Figures 3a and 3b show our measurements of the temperature profile of the regenerator and buffer tube, respectively, with no intermediate heat applied to the regenerator. With T_c = 3 K, the temperature profile is flat for almost half of the regenerator length and for about a third of the buffer tube length. The small temperature gradient at both ends of the cold heat exchanger makes the dT_m/dx power flow terms in Eq. 4 negligible at the cold end. On either side of the cold heat exchanger, we can simplify the total power flow in the regenerator $\dot{H}_{2,c}$ and in the buffer tube $\dot{H}_{2,bt}$ to

$$\dot{H}_{2,c} \approx (1 - T_c\beta_c)\dot{E}_{2,c} + \left(\frac{\epsilon_s}{1+\epsilon_s}\right)T_c\beta_c\dot{E}_{2,c} + \dot{N}h_{mol,c}, \tag{5}$$

$$\dot{H}_{2,bt} \approx \dot{E}_{2,c} + \dot{N}h_{mol,c}. \tag{6}$$

Here, we simplified the power flow expression in the buffer tube by using the condition that the oscillations in the buffer tube are nearly adiabatic ($r_h \gg \delta_\kappa \sim \delta_\nu$). A First Law analysis at the cold heat exchanger (Fig. 4) shows that the cooling power $\dot{Q}_c$ at that component is

$$\dot{Q}_c \approx \dot{E}_{2,c}T_c\beta_c\left(1 - \frac{\epsilon_s}{1+\epsilon_s}\right). \tag{7}$$

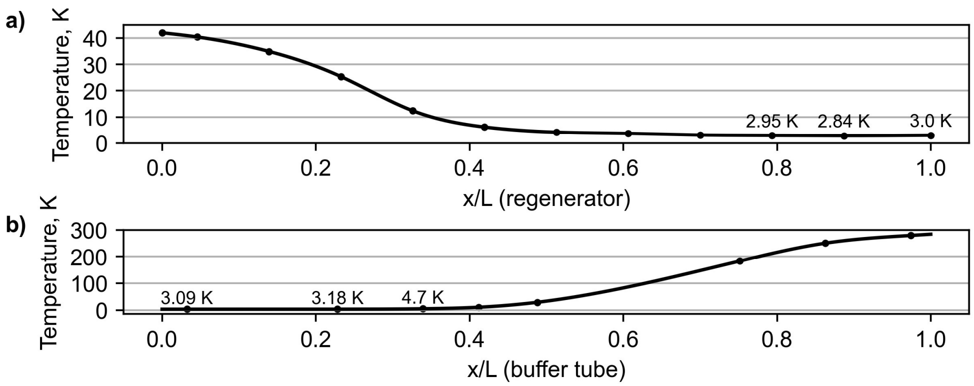


Figure 3. Temperature profile of the second-stage regenerator a) and buffer tube b) when the regenerator was regulated to 42 K at the warm end and 3 K at the cold end. Dots are thermometer measurements and lines are cubic spline fits to the measurements. No intermediate heat was applied.

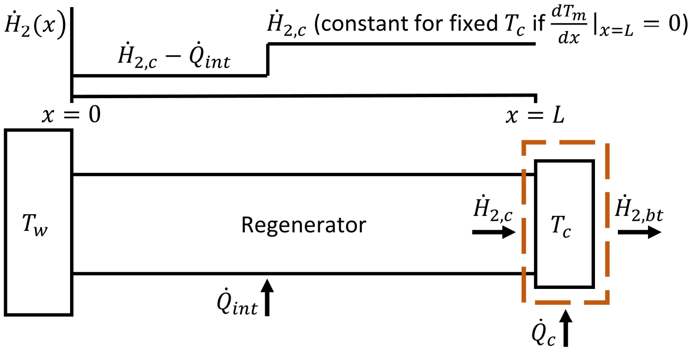


Figure 4. Schematic of the second-stage regenerator with intermediate heat applied. The top shows the idealized behavior of total power flow as a function of the distance x along the regenerator of length L. A control volume at the cold heat exchanger is used to determine that $\dot{H}_{2,c}$ is constant.

Before proceeding, we make a few additional comments on Eq. 7. As $T_c\beta_c$ approaches zero for T_c < 7 K, $\dot{Q}_c$ drops rapidly. Finite regenerator solid heat capacity ($\epsilon_s \neq 0$) adversely affects cooling power, and very low solid heat capacity drives $\dot{Q}_c$ to zero.

Changes in the cold and warm end power flows with intermediate heat

Previous work[7–11] shows that $\dot{Q}_c$ is a constant even when substantial amounts of intermediate heat are applied to the regenerator. The inset in Fig. 5a shows our measurement of this behavior in our experimental system. With no intermediate heat applied to the regenerator, $\dot{Q}_c$ of approximately 200 mW is required to reach T_c = 3 K. As the intermediate heat is swept from 0 W to 1.2 W, the $\dot{Q}_c$ required to hold T_c = 3 K is nearly constant. This observation combined with Eq. 7 requires that $\dot{E}_{2,c}$ must be constant over this range of intermediate heat inputs ($T_c\beta_c$ and ϵ_s are fixed). Using this result in Eq. 5, we conclude that $\dot{H}_{2,c}$ is also constant over this range of intermediate heat applied to the regenerator.

We now analyze the power flows around a point of intermediate heat injection using a First Law control volume around this point. Since no heat injections occurs between this point and the cold heat exchanger, the power flow leaving the control volume to the right is $\dot{H}_{2,c}$. Therefore, $\dot{H}_2$ flowing into the control volume from the left is

$$\dot{H}_{2,w} = \dot{H}_{2,c} - \dot{Q}_{int}. \tag{8}$$

No heat is injected into the regenerator to the left of the control volume, and the power flow from the injection point to the warm heat exchanger is constant and equal to $\dot{H}_{2,w}$. If we hold T_c fixed and vary $\dot{Q}_{int}$, we are making known and controlled variations to the power flow to the left of the $\dot{Q}_{int}$ injection point. $\dot{Q}_{int}$ combined with our high-resolution temperature measurements becomes a tool to study and validate the power flow equation in Eq. 4.

Figures 5a and 5b show how this tool might be used and how the restrictions discussed above are met in practice. Fig. 5a shows the point of $\dot{Q}_{int}$ injection at x/L = 0.47 (normalized distance along the regenerator axis) and measurements of the steady-state temperature profile for $\dot{Q}_{int}$ up to 1.2 W. For these experiments, T_w was fixed at 42 K and T_c was regulated to 3 K. Over the entire range of $\dot{Q}_{int}$, the temperature gradient remains near zero at the cold end and $1 - T_c\beta_c$ remains of order 1, which show that our assumptions about the power flow in the regenerator near the cold heat exchanger are valid over this range. Measurement of the temperature profile in the buffer tube (not shown) did not show appreciable changes relative to Fig. 3b, demonstrating that our

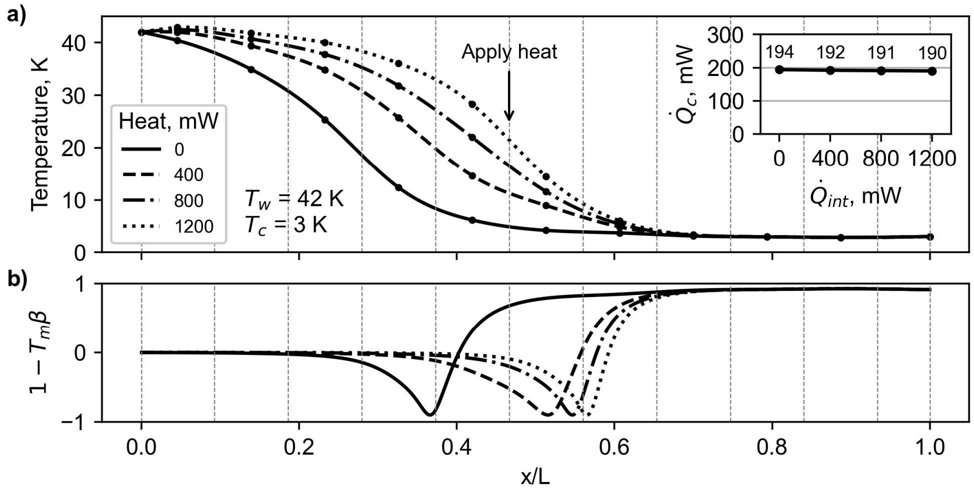


Figure 5. a) Steady state temperature profiles when different amounts of heat were applied at the location specified. Inset shows the cooling power at the cold end as a function of intermediate heat. Dotted vertical lines show all locations where heat could be applied given our system's instrumentation. b) 1-$T_m\beta$ for the same temperature profiles shown in a). x/L is the normalized length along the regenerator.

assumptions regarding power flow in buffer tube are valid over this range. A critical observation from Fig. 5a is that as $\dot{Q}_{int}$ is increased, $\dot{H}_{2,w}$ decreases (Eq. 8) and the temperature gradient at the warm end lessens so that the third and fourth terms in Eq. 4 carry less power.

A small complication at the point of intermediate heat injection

The schematic in Fig. 4 shows intermediate heat being applied to the regenerator in a step-like manner over zero distance. This is an idealization; in reality, heat must be conducted from the copper clamps to the stainless-steel regenerator tube, and then to the helium fluid and regenerator solid material. The finite heat transfer rate for all of these processes require some axial distance for $\dot{Q}_{int}$ to migrate from the perimeter of the regenerator to its core so that the power flow becomes uniform across the cross section and can be modeled with the one-dimensional equations used throughout this paper.

To estimate the axial distance ("healing length") for the power flow to become one dimensional, we applied a large amount of heat (1.5 W) at x/L = 0.65 to only one half of a copper clamp pair, leaving the other half unheated. We then recorded the temperatures of both halves of the regenerator at several upstream locations (closer to the warm heat exchanger), estimating that the power flow had returned to one dimensional once both halves were at the same temperature. In this test, the heat must migrate across an entire diameter of the regenerator. In all other experiments the heat only has to migrate from the perimeter to the center of the regenerator, so

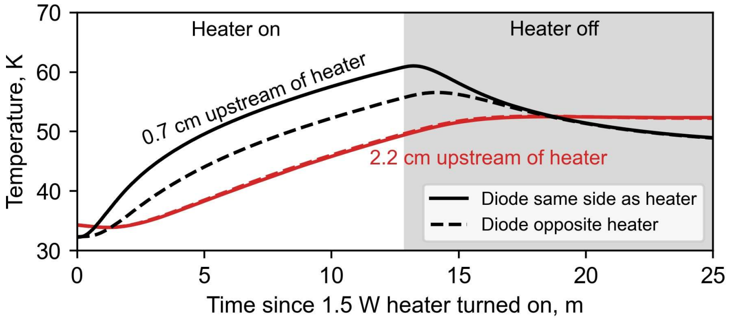


Figure 6. Evolution of temperature along the regenerator shortly after the helium compressor was turned off. At time = 0, an intermediate heat of 1.5 W was applied to one side of the regenerator to create an asymmetric heat input. Black and red lines are from diodes 0.7 cm and 2.2 cm upstream of the heater, respectively.

that this measurement results in an overestimate of the healing length. We also turned off the pressure and flow oscillations in the cryocooler, relying only on conduction and convection to transport heat across the regenerator diameter, leading to a further overestimate of the healing length.

The results are displayed in Fig. 6. At 0.7 cm upstream of the unbalanced heat injection, a significant temperature difference develops between the two regenerator halves. At 2.2 cm upstream, no temperature difference is observed. From the results of this test, it is likely acceptable to perform a one-dimensional analysis of $\dot{H}_2(x)$ at distances ≥2.2 cm upstream of the intermediate heat injection point.

Using real fluid effects to measure streaming through the regenerator

Returning to Eq. 4 and leveraging the analysis and preliminary tests described above, we describe a technique to measure the steady mass flow (streaming) through the regenerator. If sufficient $\dot{Q}_{int}$ is injected, the temperature gradient at the warm end of the regenerator can be driven to zero while maintaining a cold end temperature gradient of zero. At the warm end, the helium is close to an ideal gas so that $1 - T_m\beta \sim 0$. The regenerator solid heat capacity is large compared to the helium so that ϵ_s is small at the warm end, but because $\dot{E}_2$ is large heat capacity effects should still be accounted for. Using these results and dT_m/dx = 0 at the warm and cold ends, Eq. 8 gives

$$\begin{aligned}\left[\left(\frac{\epsilon_s}{1+\epsilon_s}\right)T_m\beta\dot{E}_2+\dot{N}h_{mol}\right]_{T_{flat}}&+\dot{Q}_{int}\\ &\approx\left[(1-T_m\beta)\dot{E}_2+\left(\frac{\epsilon_s}{1+\epsilon_s}\right)T_m\beta\dot{E}_2+\dot{N}h_{mol}\right]_{T_c},\end{aligned} \tag{9}$$

where T_{flat} is the temperature near T_w where the gradient is zero. Equation 9 can be rearranged into an expression for the molar steady flow

$$\dot{N}\approx\frac{\left[\left(1-\frac{T_m\beta}{1+\epsilon_s}\right)\dot{E}_2\right]_{T_c}-\left[\left(\frac{\epsilon_s}{1+\epsilon_s}\right)T_m\beta\dot{E}_2\right]_{T_{flat}}-\dot{Q}_{int}}{h_{mol,flat}-h_{mol,c}}. \tag{10}$$

The measurements in Fig. 7 show that by injecting $\dot{Q}_{int}$ = 1.52 W near the middle of the regenerator, we can force the temperature gradient at the warm end to zero. The acoustic power in Eq. 10 can be estimated using Eq. 7 and the measured $\dot{Q}_c$. The remaining variables can be computed from the known properties of helium and the regenerator solid. At present we do not know the regenerator solid materials used in this commercial cryocooler; however, if the material

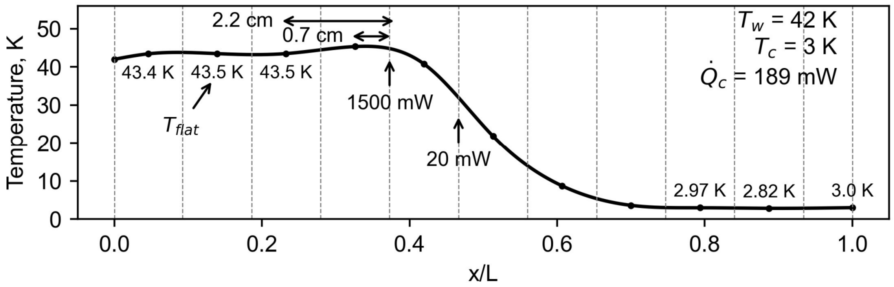


Figure 7. Regenerator temperature profile when 1.52 W of heat was applied to the locations shown by the arrows. Note that the temperature 0.7 cm upstream of $\dot{Q}_{int}$ application is higher than the temperature 2.2 cm upstream. This is possibly an effect of underdeveloped $\dot{H}_2$, as discussed in the previous section. T_{flat} is the temperature near the warm end where the gradient is zero.

at the cold end is $HoCu_2$ and the material at the warm end is Pb, then the streaming flow is 0.8 mmol/s.

Isolating temperature and temperature gradient terms of the total power equation

The experiments in Fig. 8 show how we can leverage and extend the results and techniques described above to examine individual terms in the total power equation. Here, we applied intermediate heat at two distinct locations labeled $\dot{Q}_{sweep}$ and $\dot{Q}_{control}$. Extending the First Law analysis from Fig. 4 to this setting, $\dot{Q}_{sweep}$ is used to vary the power flow by a known amount at the upstream location labeled "Measurement Plane". Further upstream, $\dot{Q}_{control}$ is used to adjust the temperature profile and fix either T_m or dT_m/dx at the Measurement Plane. The combination of these two intermediate heat injections gives the sensitivity of T_m or dT_m/dx to changes in power flow, which can be used to validate these terms in the power flow expression in Eq. 4.

In the specific experiment shown in Figs. 8a and 8b, the measurement plane is at $x/L \sim 0.6$, and we use $\dot{Q}_{control}$ to fix T_m at 7.5 K. $\dot{Q}_{sweep}$ is varied from 0 mW to 350 mW, and dT_m/dx is computed at the measurement plane using spline fits to the temperature measurements. The accuracy of the temperature control at the measurement plane is displayed in the Fig. 8a inset and in the middle plot in Fig. 8b. The measurements of $\dot{Q}_c$ in bottom plot of Fig 8b and the temperature profiles near the cold end in Fig 8a show that this experiment stays within the range of validity discussed above, i.e. that $\dot{H}_{2,c}$ is constant.

The change in power flow at the measurement plane is the negative of the heat injection at the sweep injection point, i.e., $\Delta\dot{H}_2 = -\dot{Q}_{sweep}$. Since T_m is fixed at the measurement plane, the first two terms on the right hand side of Eq. 4 cannot adjust to generate $\Delta\dot{H}_2$: all of the variables in these terms are only a function of T_m, including $\dot{E}_2$. The fifth term on the right hand side of Eq. 4 is the

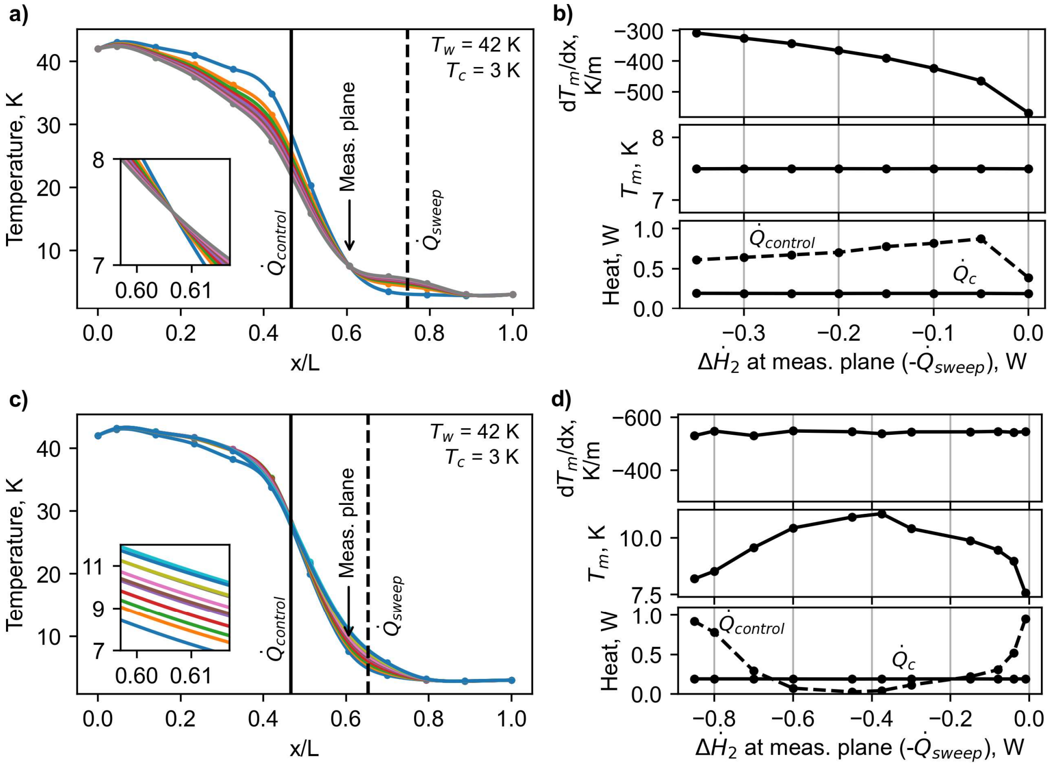


Figure 8. a) Temperature profiles of the regenerator when different amounts of sweeping heat were applied at the location of the vertical dashed line. Control heat was applied at the solid vertical line to keep the temperature at the measurement plane constant. Inset shows a zoom of the measurement plane. The temperature, temperature gradient, cooling power, and control heat for a) are shown as functions of the sweeping heat in b). c) and d) are similar except that temperature gradient was controlled.

power flow from steady streaming. This term does not adjust to generate $\Delta\dot{H}_2$ because the specific molar enthalpy is fixed by T_m and the streaming flow ($\dot{N}$) is not significantly affected. The only remaining terms in Eq. 4 that can adjust are those proportional to dT_m/dx. Solid and fluid conduction in porous media regenerators are often negligible compared to the convection term in Eq. 4 (third term on the right hand side). Therefore, the power flow change at the measurement plane is approximately given by

$$-\dot{Q}_{sweep} = \Delta\dot{H}_2 \approx F(x,T_m)\,\Delta\left\{|U_1|^2\frac{dT_m}{dx}\right\}, \tag{11}$$

where $F(x,T_m)$ has been pulled outside of the $\Delta\{\ldots\}$ because $F(x,T_m)$ should be constant for fixed T_m (see Eq. 3). We nominally expect $\dot{Q}_{sweep}$ to be proportional to the change in dT_m/dx at the measurement plane. The top plot in Fig. 8b shows the trend for dT_m/dx versus $-\dot{Q}_{sweep}$. The general trend matches our expectation that reductions in $\dot{H}_2$ at the Measurement Plane results in a reduction in the magnitude of dT_m/dx. The relationship is relatively linear after the first reduction in total power, i.e. for $\Delta\dot{H}_2 < -0.05$ W.

We have considered two potential explanations for the nonlinear dependence between $\Delta\dot{H}_2$ and dT_m/dx. The first is related to the rapid changes in the compressibility of helium between 6 K and 8 K as it moves from liquid-like to ideal gas-like behavior. For the initial step in $\dot{Q}_{sweep}$, the helium between the measurement plane and the cold heat exchanger is in the liquid-like domain with very low compressibility, which leads to lower values of $|U_1|^2$at the measurements plane and higher changes in dT_m/dx per $\Delta\dot{H}_2$. For the subsequent steps in $\dot{Q}_{sweep}$, the helium between the measurement plane and the cold heat exchanger is more toward the ideal gas compressibility, which leads to higher values of $|U_1|^2$at the measurements plane and lower changes in dT_m/dx per $\Delta\dot{H}_2$. The second potential explanation of the nonlinear dependence between $\Delta\dot{H}_2$ and dT_m/dx is a transition to or from temperature asymmetry. As discussed in our group's companion paper at this conference,[12] large temperature asymmetries may exist between the two halves of the regenerator. In that case, it is not appropriate to assume $T_m(x)$ is azimuthally uniform, and the traditional thermoacoustic simplification of $\dot{H}_2(x)$ being one-dimensional no longer holds. In all future work we will attempt to restrict analyses like these to temperature profiles that we know are azimuthally symmetric, perhaps by influencing the profile with asymmetric intermediate heat.

Using the same methodology displayed in Figs. 8a and 8b, we have also controlled for constant dT_m/dx at the measurement plane while varying $\dot{Q}_{sweep}$. These experiments were performed before our improved understanding of the "healing length" requirement discussed above, and $\dot{Q}_{sweep}$ was added just 0.7 cm from the measurement plane – too close to ensure the power flow is one dimensional. However, Figs. 8c and 8d demonstrate the feasibility of this technique and our ability to isolate the variation of T_m with $\Delta\dot{H}_2$ at the Measurement Plane.

CONCLUSION

Real fluid properties affect low-temperature cryocoolers in several important ways. As has been reported previously[7–11] and as we have found in our experiments, a sizable amount of intermediate heat can be applied to low-temperature regenerators without affecting cooling power at the cold heat exchanger (at least 1.52 W as shown by our study). The ability of the regenerator to absorb and transport this intermediate heat is derived from the flexibility in the temperature profile between the point of injection and the warm heat exchanger. Without intermediate heat applied, the temperature profile near the cold heat exchanger is flat and the temperature gradient is concentrated near the warm end. As intermediate heat is applied, the temperature gradient adjusts to changes in total power and, on average, becomes less steep.

Through analysis of the total power equation and measurements of the temperature profile, we showed that the power flow at the cold end is unaffected by injection of intermediate heat along the axis of the regenerator because the power flow into the cold heat exchanger is fixed by real fluid properties. The injected heat only modifies the power flow in the region between the injection point and the warm heat exchanger. This conclusion is dependent on the low-temperature regenerator operating within a few bounds of validity: the cold end temperature stays fixed, the temperature gradient remains nearly zero on both the regenerator and buffer tube sides of the cold heat exchanger, and any steady streaming flow through the regenerator stays nearly constant. Our experiments demonstrate that these conditions are met for a wide range of intermediate heat injections.

Leveraging these observations, we developed several powerful techniques that use changes in intermediate heat input to isolate and measure individual terms in the power flow equation. Using a single intermediate heat injection, we forced the temperature gradient to zero at the warm heat exchanger and measured the power flow carried by steady streaming flow through the regenerator. Using two, spatially separated intermediate heat injections we demonstrated techniques to isolate and measure individual terms in the power flow equations, i.e., terms dependent on either dT_m/dx or T_m. In future work, we will improve the analysis and measurement presented here to validate the terms in the power equation – a particularly important goal, as cryocooler performance is inextricably tied to this equation.

ACKNOWLEDGMENTS

R. Snodgrass acknowledges support from a National Research Council Postdoctoral Fellowship. We also thank Greg Swift for stimulating discussion and leadership, and for assistance with the thermoacoustic software DeltaEC. We thank Ray Radebaugh for guidance during preliminary project planning, and Nathan Ortiz for experimental support.

Figures were generated using Matplotlib and material properties were gathered using CoolProp[13], an open-source thermophysical library available as a Python wrapper.

REFERENCES

1. de Waele, A.T.A.M., Xu, M.Y., and Ju, Y.L., “Nonideal-gas effect in regenerators,” *Cryogenics*, vol. 39, no. 10, (1999), pp. 847–851.
2. Cao, Q., Qiu, L., and Gan, Z., “Real gas effects on the temperature profile of regenerators,” *Cryogenics*, vol. 61, (2014), pp. 31–37.
3. Lang, A., Häfner, H.-U., and Heiden, C., “Systematic Investigations of Regenerators for 4.2K-Refrigerators” in *Advances in Cryogenic Engineering*, Boston, MA: Springer US, pp. 1573–1580, (1998).
4. Rott, N., “Thermoacoustics” in *Advances in Applied Mechanics* vol. 20, Elsevier, pp. 135–175, (1980).
5. Swift, G.W., *Thermoacoustics: A Unifying Perspective for Some Engines and Refrigerators*, 2nd ed. Springer International Publishing, (2017).
6. Ward, B., Clark, J., and Swift, G., “Users Guide for DeltaEC: Design Environment for Low-amplitude Thermoacoustic Energy Conversion.” (2017), [Online]. Available: www.lanl.gov/thermoacoustics.
7. Ravex, A., Trollier, T., Tanchon, J., and Prouvé, T., “Free Third-Stage Cooling for Two-Stage 4 K Pulse Tube Cryocooler,” in *Cryocoolers 14*, pp. 157–161, (2007).
8. Wang, C., “Extracting Cooling from the Pulse Tube and Regenerator in a 4 K Pulse Tube Cryocooler,” in *Cryocoolers 15*, pp. 177–184, (2009).
9. Prouvé, T., Godfrin, H., Gianèse, C., Triqueneaux, S., and Ravex, A., “Experimental results on the free cooling power available on 4K pulse tube coolers,” *J. Phys. Conf. Ser.*, vol. 150, no. 1, (2009), p. 012038.
10. Zhu, S., Ichikawa, M., Nogawa, M., and Inoue, T., “4 K pulse tube refrigerator and excess cooling power,” *AIP Conf. Proc.*, vol. 613, no. 1, (2002), pp. 633–640.
11. Uhlig, K., “3He/4He dilution refrigerator with high cooling capacity and direct pulse tube pre-cooling,” *Cryogenics*, vol. 48, no. 11, (2008), pp. 511–514.
12. Snodgrass, R., Kotsubo, V., Ullom, J., and Backhaus, S., “A Temperature Instability in 4 K Cryocooler Regenerators Caused by Real Fluid Properties,” *Cryocoolers 21*, (2021).
13. Bell, I.H., Wronski, J., Quoilin, S., and Lemort, V., “Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp,” *Ind. Eng. Chem. Res.*, vol. 53, no. 6, (2014), pp. 2498–2508.

**Proceedings of the ASME 2021 16th International
Manufacturing Science and Engineering Conference
MSEC2021
June 21-25, 2021, Cincinnati, OH, USA**

MSEC2021-1957

TOWARDS A DIGITAL DEPOT TO SUPPORT SUSTAINABLE MANUFACTURING DURING CRISIS RESPONSE

Nancy Diaz-Elsayed[1]
Smart and Sustainable Systems Laboratory
Department of Mechanical Engineering
University of South Florida
Tampa, FL, USA

KC Morris
Systems Integration Division
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD, USA

Julius Schoop
Department of Mechanical Engineering and the
Institute for Sustainable Manufacturing
University of Kentucky
Lexington, KY, USA

ABSTRACT

The COVID-19 pandemic has imposed new challenges to maintaining sustainability in our manufacturing operations. With such high variability in demand for urgently needed products (e.g., personal protective equipment, testing technologies) and shifts in the needed capabilities of already complex production systems, sustainability challenges concerning waste management, life cycle impact characterization, and production operations have come to light. An extensive amount of data can be extracted from manufacturing systems, but it is not yet being used to improve the performance of production systems and maintain sustainability strategies during times of distress. This article proposes the concept of a Digital Depot. Being virtual in nature, the depot can contain plans and data for many different types of crises and contain a wider array of products than would be available in a physical, national stockpile. The information could be made available on demand to a national base of manufacturers to help them swiftly pivot to the production of critically needed goods while building on their existing manufacturing capabilities. The contents of the Digital Depot would be applicable to several stages pertinent to manufacturing operations including product definition, production planning information, asset and factory-level data, as well as data concerning the supply chain, distribution, and end-of-life stages. Future work is recommended in the development of templates for robust and secure data sharing, as well as multi-disciplinary identification of businesses cases for data-driven collaborative and sustainable manufacturing practices enabled by the Digital Depot.

Keywords: Smart Manufacturing; Manufacturing Ecosystem; Supply Chain; Sustainable Manufacturing; Resiliency; Digitization; National Stockpile; Manufacturing on Demand; Life Cycle Assessment; Life Cycle Inventory

1. INTRODUCTION

The COVID-19 pandemic ushered in new challenges for the manufacturing industry including sharp rises and steep declines in consumer demand while coping with a volatile supply chain [1–3]. Reactions to these changes happened swiftly— a necessity in an effort to maintain human life. Nonetheless, such variability in demand requires exceptional agility and resources to allow manufacturers to swiftly pivot to the production of critical goods. Without the right practices and key engineering artifacts in place, the outcome could naturally lead to waste (in a variety of forms such as time and materials) and augment the negative impact that society and the manufacturing sector has on our environment. The COVID-19 pandemic highlighted that improvements were indeed needed to the collection and maintenance of our nation's emergency response stockpiles, which was further emphasized by President Joseph Biden Jr.'s Executive Order for more resilient supply chains [4].

An Atlantic article cites Health and Human Services Secretary Alex Azar as reporting the stockpile contained only 42 million face masks when 3.5 billion of these masks were needed [5]. The

[1] Contact author: nancyd1@usf.edu

Morris, KC; Diaz-Elsayed, Nancy; Schoop, Julius. "Towards a Digital Depot to Support Sustainable Manufacturing During Crisis Response."
Presented at ASME Manufacturing Science and Engineering Conference 2021 (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021.

article raises two shortcomings: the understocking of the stockpile and the onerous regulations on the production of supplies. Supplying a nation of over 331 million people on short notice would be no small feat, not to mention the demand of 7.8 billion people across the globe. Manufacturers have recognized this challenge as they have shifted their prioritization of Industry 4.0 objectives towards agility in operations and flexibility in the customization of products since the onset of the pandemic [6]. Furthermore, sidestepping regulations on this type of equipment that is meant to protect human health and safety would come with risks. Technical solutions that can help speed the testing and certification of products are needed.

In this paper, the concept of a Digital Depot is proposed to help address any forthcoming emergency response needs. Under this concept, product designs and manufacturing instructions can be stored along with methods for vetting suppliers that will be able to create new products on demand in response to emergency needs. In addition, sustainability aspects of the mass production of new products can be assessed and planned for before production begins. The Digital Depot can build on current trends in manufacturing including advanced assessment and planning for environmental impacts, manufacturing-as-a-service, digital certification of production, and virtual training for manufacturing and maintaining unique products. The Digital Depot offers an opportunity to reduce physical stockpiles of fully assembled products, and instead allows for a pivot towards 'stockpiling data' and raw materials, which could be used to produce a variety of relevant products during specific crisis situations. In this paper, we describe the concept of a Digital Depot in light of these trends and discuss how the approach can be used to reduce environmental risks for future large-scale response efforts. A policy component would need to be developed to fully deploy the technical capabilities outlined.

2. INCORPORATING SUSTAINABILITY INTO THE DIGITAL DEPOT

A Digital Depot would consist of a comprehensive set of planning tools focused on producing the goods that are needed to respond to national crises. Being virtual in nature, the depot can contain plans and data for many different types of crises and contain a wider array of products than would be available in a physical stockpile. The key concept is that these plans could be made available on demand to a national base of manufacturers to help them to swiftly pivot to manufacture new types of goods building on their existing manufacturing capabilities. Data for the Digital Depot can be acquired from existing databases and systems already being utilized by many manufacturers (e.g., Enterprise Resource Planning [ERP], Materials Resource Planning [MRP], Certified Reference Materials [CRM]) through the development of a digital thread for the critical products necessary for crisis response. By taking the time for advanced planning we should be able to anticipate more scenarios for production. Considerations from the basic supplies that may be needed to physical stockpiles of appropriate raw materials to end-of-life plans for rapidly surging products can be incorporated into plans.

In Diaz-Elsayed et al. [2], the authors identified three key risk areas for unanticipated consequences of the rapid pivoting to new products and supply chains that could have negative impacts on our environment and workforce:

- **New modes of operation for manufacturing**: The vast majority of job functions under COVID-imposed conditions have shifted. Many jobs, particularly in the manufacturing sector, pivoted to remote work when possible [7]. Functions from management and manufacturing engineering to training and maintenance all found a remarkable capacity for remote work. Others operated under alternative and extended schedules in order to limit the number of people on site at any given time.
- **Life cycle impacts of new and fluctuating product streams**: The environmental impacts of manufacturing products are highly dependent on the material inputs and manufacturing processes. The complexity of assessing life cycle environmental impacts using current methods means that these assessments are not readily available during planning phases of product design and production. In addition, supply chain disruptions can also result in poor material or inefficient process choices. The result will be poor balancing of benefits with environmental impacts.
- **Increases in waste generation**: Waste created by rapid changes in supply and demand during the COVID-19 pandemic sky-rocketed, and concerns for the dangers raised to the environment naturally followed [8–10]. Hospitals reported spikes in medical waste and supply chains of recycled materials destined for manufacturers were disrupted by the changing consumption patterns [11,12]. Producers of consumer and commercial goods worked hard to find different outlets for the products and pivot to different supply chains with mixed success. Dairies and other food industries reported significant waste increases of their products brought on by the supply chain disruptions [13].

This paper describes how these risk areas can be addressed by leaning on the Digital Depot concept to help quantify and plan for pending impacts.

3. WHAT IS THE DIGITAL DEPOT?

A central function of the Digital Depot would be the 'virtual stockpiling' of relevant product, production plans, and data for components and assemblies needed during potential emergencies. Rather than stockpiling the physical (manufactured) goods in advance, which may result in either overproduction or long-term decay and malfunctions of such goods, the Depot would enable rapid on-demand production by supplying relevant digital files to previously vetted manufacturers (see Figure 1). Such data, including Computer

Aided Design (CAD) and Computer Aided Manufacturing (CAM) files, would necessarily go beyond simple geometric definitions of components, and encompass ‘design intent’ to allow local engineering and production teams to produce each component for ‘fitness in service,’ rather than to simply meet the dimensions of an annotated 3D CAD file. Indeed, connecting the product design and performance (use stage) with the manufacturing processes used to create these processes is a key challenge the Digital Depot will need to address in order to yield actionable results and true resilience in light of future crisis situations. In addition, the Depot contains guidance for rapid formation of business partnerships to facilitate rapid creation of new supply chains.

The following key components of a Digital Depot will be instrumental in incorporating sustainability thinking into the overall planning for *manufacturing for crisis response*:

- List of covered products and their performance requirements
- Digital designs for the products and components
- Mapping of component and supply chains needs
- Production instructions and training materials for unique components
- Instructions for certification of products against performance requirements
- Assessment of sustainability impacts of digital designs
- Assessment of sustainability impacts of production
- Registry of vetted suppliers for product components and assemblies
- Plans for end-of-life reclamation of identified products and by-products of their production

The Digital Depot as a concept includes two types of data distinguished by their temporal criticality. Fundamentally, the Depot is a repository or registry of the plans for producing a wide array of products needed to respond to a crisis situation. However, crisis response also demands access to temporally sensitive data such as asset and supply chain capabilities and availability that can support real-time planning. This paper limits discussion to the concept of these capabilities and does not address how they might be deployed in the real world. Understandably, access to real-time data would introduce a broad set of concerns ranging from cyber security to protection of privacy and intellectual property. Similarly, a balance between top-down planning of the system versus a demand response model of implementation will be needed to optimize efficiency and encourage innovation. Likewise, the architecture of the depot need not be monolithic, but it could be a composition of independently offered services. The implementation of the Digital Depot concept would require much thought and discussion to find a suitable balance around such factors. The Digital Depot described here represents the concept as an umbrella for envisioning a future manufacturing paradigm to support crisis response.

Figure 1 illustrates the functional areas needed to support a rapid on-demand manufacturing response to changing demands under crisis in a sustainable manner. Central to that response is the Digital Depot that contains product and production information required to activate the response. Each area is described below.

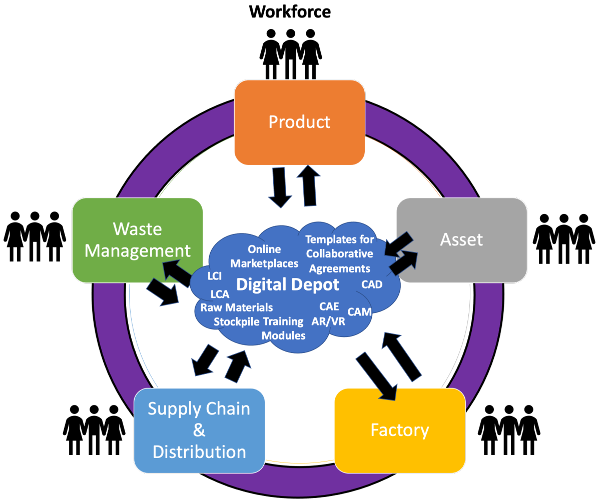


Figure 1: Examples of content that may be included in the Digital Depot with respect to production and product life cycle stages. Abbreviations: AR: Augmented Reality; CAD: Computer Aided Design; CAE: Computer Aided Engineering; CAM: Computer Aided Manufacturing; LCA: Life Cycle Assessment; LCI: Life Cycle Inventory; VR: Virtual Reality.

Product: The incorporation of product data into the Digital Depot serves two purposes. First, with advance planning of supplies, the products, materials, and production alternatives can be evaluated for their impacts across all three risk factors (waste management, life cycle impacts, and new modes of operations). Designs for the mass production of items can be pre-evaluated for life cycle properties such as material choice and waste generation. The range of products and production processes contained in the depot can include low technology items such as face masks that are needed in large quantities to very sophisticated items such as high-tech ventilators or equipment for vaccine production. By vetting the designs available, the impacts of the products in terms of waste generation and product reclamation at the end-of-life stage can be considered. Secondly, societal needs can be anticipated should the product end up being produced. We can start advance planning for the disposal systems that will be needed as well as plan for training on how the products are produced and used. In addition, virtual stockpiling mechanisms can be developed to avoid stockpiling fully assembled physical products. The Digital Depot enables rapid and agile production of such products on-demand during a crisis situation.

Asset: At the asset level, the Digital Depot can store life cycle inventory data for conventional and new products to facilitate decision-making for process planning for pending orders. Relevant life cycle inventory data for the asset include resources consumed (e.g., energy, water, materials) and are discussed further in Section 4.1. Although data analytics are outside of the scope of the functions of the Digital Depot, such services could be provided external to the depot to facilitate the sharing of production capabilities across manufacturers to expand capacity and meet large variations in demand and to propose enhanced process plans to lower the costs and environmental impacts of the asset.

Factory: At the factory level, the optimization of production operations (e.g., to reduce carbon emissions and energy consumption) can be supported by the availability of design data and real-time information. CAD files of the production equipment within a factory (existing equipment and/or planned purchases) can be used to optimize the layout of a factory and investigate strategies to reduce waste, such as wait and transportation time, in a production line. As a factory strives to improve their performance, the data acquired can support the validation of optimization strategies that have been implemented to further enable operational improvements. From a business planning perspective, the classification of a factory's digital readiness can provide management with insight about the facility's ability to adopt new digital technologies and make advancements towards their Industry 4.0 roadmap [14–16]. Additionally, a factory's plan for continuity of operations (e.g., [17]) can provide customers with insight about the facility's reliability during times of emergency.

Workforce: To support an agile and digitally-enabled workforce, the Digital Depot could provide relevant training modules (e.g., videos and digital study materials) related to specific manufacturing process and assembly tasks. For example, such materials could focus on enabling transfer learning between a given company's current workforce activities and similar, yet sufficiently different, skills required to produce a new product (e.g., a ventilator) on-demand. Instead of relying on traditional (in-person) modalities, the goal of digital workforce training modules provided through the Depot would be to allow for rapid up-skilling of the incumbent workforce. Such training would also encompass sustainability considerations, which require not just advanced technologies, but also specific workforce behaviors and skills.

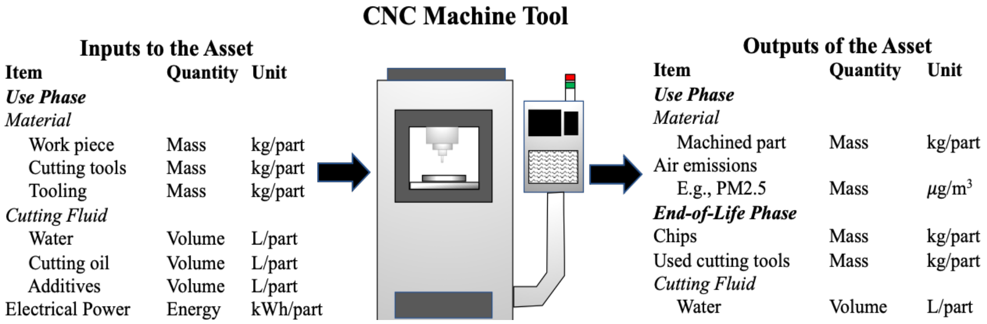
Supply Chain and Distribution: A major lesson of the global pandemic has been humanity's reliance on supply chains and their rigidity in light of changing patterns of consumption. The disruption of supply chains has resulted in much waste, a significant burden on the workforce, and exceptional stress levels across society. On the flip-side technologies such as digital marketplaces enabled reconfiguration of supply chains and insights into changing consumption patterns. More pervasive use of these technologies will enable better response to future crises. Similarly, organizations that specialize in distribution, such as UPS[2], UberEats and similar services have been essential to the restructuring of supply chains and delivery to consumers Supply chains are also seeing a convergence towards leveraging shipping services [18]. The efficiency introduced through the specialization of distribution modes has the potential to result in reduced environmental impacts as the use of transportation is optimized.

Waste Management Networks: The COVID-19 impacts on waste management are only starting to be understood, but early results indicate their existence [8,10,19]. Efforts towards fostering a circular economy are growing rapidly and require much more sophisticated waste management than currently exists. An infrastructure around material identification and repurposing methods is needed. Considerations for design for remanufacturing of emergency supplies can be complimented with plans for recycling of materials through increased variety in waste management capabilities combined with specialized collection methods. In other words, digital marking of materials can help direct waste to specialized processing facilities, whether those be remanufacturing facilities or specialized material reclamation facilities. Specialized distribution networks can be established in for these purposes as well as product distribution.

Underpinning the Digital Depot are standard specifications for sharing product and manufacturing process data to be used across business partners. Standards in this area are rapidly evolving to support virtualization of manufacturing systems and services [20]. The Digital Depot will rely on the following standards for product and production systems design incorporating sustainability assessment. (This list is not exhaustive; certainly other standards will also be critical to the design and operation of a Digital Depot.)

- Product data definitions standards are now successfully deployed for sharing product designs and associated information [21,22].
- ISO 23247 is now emerging with a Digital Twin Framework for Manufacturing [23].
- Standards for measuring the sustainability impact of products [24] have been successfully used for obtaining rough estimates of impacts across a range of products, with an ever-growing number of product category definitions [25] emerging as well as more and more refined data reflecting the life cycle impacts of component materials and services.
- Standards for sustainability assessments of manufacturing processes and production systems have been recently released from ASTM International [26,27] and ISO [28,29].

[2] Identification of commercial products or services here does not imply recommendation nor endorsement by NIST, nor is it intended to imply that they are necessarily the best available for the purpose.



Inputs to the Asset

Item	Quantity	Unit
Use Phase		
Material		
Work piece	Mass	kg/part
Cutting tools	Mass	kg/part
Tooling	Mass	kg/part
Cutting Fluid		
Water	Volume	L/part
Cutting oil	Volume	L/part
Additives	Volume	L/part
Electrical Power	Energy	kWh/part

Outputs of the Asset

Item	Quantity	Unit
Use Phase		
Material		
Machined part	Mass	kg/part
Air emissions		
E.g., PM2.5	Mass	μg/m^3
End-of-Life Phase		
Chips	Mass	kg/part
Used cutting tools	Mass	kg/part
Cutting Fluid		
Water	Volume	L/part

Figure 2: Examples of the types of inputs and outputs (inventory data) that would be collected for a CNC machining center for a life cycle assessment. CNC: Computer Numerical Control.

4. TRENDS IN MANUFACTURING RESEARCH TO ADDRESS SUSTAINABILITY RISKS

The framework for a Digital Depot for emergency response efforts will be a massive undertaking. Building sustainability considerations from the start is imperative to control sustainability impacts. The section highlights two key areas where today's research and development can be leveraged to address sustainability concerns in the planning effort:

- Advanced assessment and planning for environmental impacts, and
- Capabilities for manufacturing on demand.

4.1 Informing Environmental Impact Assessments

A Life Cycle Assessment (LCA) is often used to evaluate the environmental impacts (e.g., greenhouse gas [GHG] emissions) of a product, process, or system. The steps of an LCA consist of 1) Goal and Scope Definition, 2) Inventory Analysis, 3) Impact Assessment, and 4) Interpretation [24]. As the life cycle impacts of manufacturing processes vary greatly depending on category and even the model of production equipment being used, estimates of life cycle environmental impacts are expected to have significant variation and uncertainty.

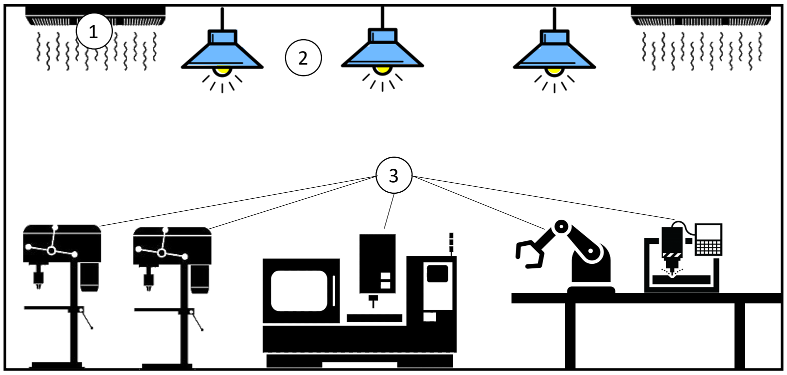
The COVID-19 pandemic has introduced many challenges to the manufacturing sector, including spikes in demand for personal protective equipment (PPE) for the healthcare industry and broader public. The variation in demand for PPE has been reflected in online portals such as Project N95 and Thomas [2]. Table 1 presents Greenhouse gas (GHG) emissions for alternative types of masks. The table shows that depending on the assumptions made, the variance in GHG emissions can be significant (e.g., by modifying the type of material being used, introducing transportation impacts, or changing the number of washes for a cloth mask).

This variation in emissions generation in mask production, suggests other sustainability metrics will also vary. Similarly, but not captured here, one can expect significant variation in impacts from the type of production processes used in mask production [30] or even different design types with a single process [31].

Table 1: Greenhouse gas (GHG) emissions for the production of personal protective equipment (PPE).

Product	GHG Emissions	Notes	References
Impacts per Part			
N95 Mask	0.05 kg CO2-eq/part (single use)	Excludes transportation; Major impact: PP material	[32]
Surgical Mask	0.059 kg CO2-eq/part (single use)	Includes transportation; Major impact: Transportation	[8]
Cloth Mask	0.06 kg CO2-eq/part (single use)	Excludes washing; Major impact: Cotton fabric	[32]
Cloth Mask	6.92 kg CO2-eq/part 0.036 kg CO2-eq/usage	Includes washing; assumes mask is used 183 times	[8]
Impacts by Filter Efficiency			
Cloth Mask	0.0072 kg CO2-eq/filter efficiency	Assumes ~50% filter efficiency	[33]
Surgical Mask	0.0074 kg CO2-eq/filter efficiency	Assumes ~80% filter efficiency	[33]

With respect to conducting the LCA, there are three dominant types of methods used: process, economic input-output life cycle assessment (EIO-LCA) or a hybrid approach that couples these two methods. For a process LCA, the methodology is known to be data and time intensive, which can be expensive. Nonetheless, a process LCA provides better accuracy of the life cycle environmental impacts for specific products, systems, and/or processes being evaluated when compared to the amortized impacts available for the production of products within specific industries as is done with EIO-LCA.



(1) HVAC System:
- Electricity demand
- Filter replacement intervals
- Refrigerant emissions

(2) Lighting:
- Electricity demand
- Bulb replacement intervals

(3) Production Equipment:
- Electricity demand
- Process data (e.g., process parameters)
- Performance data (e.g., process and setup time)
- Water demand

Figure 3: Factory-level information relevant to the life cycle assessment. HVAC: heating, ventilation, and air conditioning.

As a process LCA is very data intensive, the availability of data on the production floor would serve to inform the Inventory Assessment of the LCA. For example, the inputs and outputs of a machining process are depicted in Figure 2 to provide exemplary data of the type of data that would be collected for the Inventory Assessment phase of the LCA. The power demand is known to vary with time during the production of a part. Nonetheless, models have been developed to predict the energy consumed during the machining process [34]. Other relevant resources that are consumed include cutting fluids, workpiece materials, and cutting tools. While the cutting fluid consumed can be amortized over the parts produced, some process data for the LCA can be extracted from a CAD file (e.g., the weight of the material being used for the work piece).

In order to holistically obtain life cycle data for the production of a part, factory resource consumption should be considered as well [35]. Moreover, local factors, such as the climate, should be accounted for as they can influence the resources consumed by the factory [36]. Important inputs at the factory level include energy to power the lighting and heating, ventilation, and air conditioning systems and water consumption (see Figure 3). Relevant data on resource use can be extracted via conventional methods, such as utility bills. Alternative methods such as metering can provide more granular information about how the resources are consumed to more accurately attribute the resources consumed to specific production processes and parts. Meter readings could be integrated with an energy management application that stores historical data at a finer scale than what is available through a utility bill, and would grant visibility to the consumption of resources at the sub-system or asset level.

4.2 Enabling Manufacturing on Demand

The Digital Depot should contain enough information to enable manufacturing of products on demand, which is key to enabling an alternative to the current paradigm of stockpiling vast quantities of fully assembled products that may be subject to finite shelf life or experience degraded performance over time. As describe in the following section, in addition to product designs and production plans, also necessary is information on design intent, connections for integrating with enterprise and e-commerce data, cybersecurity preparedness, and support for workforce readiness.

The proposed concept of a Digital Depot is particularly powerful and relevant to the manufacturing stage for products that are complicated to produce (i.e., the majority of modern products). Even seemingly ‘simple’ items, such as N95 masks and replacement parts for medical equipment, contain a considerable amount of non-trivial features and manufacturing steps in their production. For example, handling and processing of filter fabric, including joining (gluing) dissimilar materials is paramount to realizing required product performance (e.g., N95 level of filtration and durability) [37]. Traditional blueprints and even modern 3D CAD drawings fail to fully describe key functional features and design intent, which may lead to unacceptable product quality. The importance of process-induced characteristics, such as machining-induced surface integrity, on functional performance cannot be overstated, and requires strong ‘physics-based’ understanding of unit processes [38]. Based on this realization, several efforts to include in more than just simple geometric information in the digital definition of a part have therefore been undertaken [39]. While such model-based feature identification is clearly needed to produce highly engineered components such as turbine blades and biomedical implants, the importance of this information for simpler components has become evident during the COVID-19 pandemic.

In addition to ‘design intent,’ we also propose that important sustainability consideration related to both production and use-stage performance be included in the ‘data package’ provided by the Digital Depot. Such data could include waste and environmental impact considerations, including assembly and distribution/storage information. In this manner, supply chain-wide insights regarding material, waste and energy flows could be promoted more proactively, which is envisioned as the foundation for more environmentally benign manufacturing operations.

One of the key hurdles to realizing an agile and 'real-world ready' Digital Depot lies in a 'black box' approach towards manufacturing processes that neglects the physics that govern these production steps. Manufacturing is inherently complex and multi-disciplinary, as the laws that govern even simple processes are often highly non-linear, and thus non-intuitive. By leveraging the concept of the Digital Thread and Digital Twin of manufactured components, the Digital Depot paradigm could enable agile production, while also allowing for key sustainability metrics to be considered. In contrast to traditional stockpiling and warehousing of fully assembled and manufactured goods, a Digital Depot would merely require key raw materials to be stored in limited quantities. Centrally stored digital component and assembly definitions, including material properties, manufacturing process information, and even end-of-life considerations (recoverability, reusability, recyclability) could be made available to the industrial base at short notice, in order to quickly pivot production to new product lines needed to address a crisis situation. Since the Digital Depot would embody billions of dollars of intellectual property (IP), careful consideration of cybersecurity concerns will be vital to its successful implementation.

To ensure cybersecurity and 'need to know' access to key Digital Depot information, virtual certification of participating manufacturers is proposed. Such certifications could include training and process-focused, generic 'first article' demonstration of a manufacturer's ability to effectively work with a data set provided by the Digital Depot. Since it will not be possible to anticipate every possible kind of product or component that may be required in an emergency situation, certification of manufacturers according to specific processes (filter fabric production, precision machining, etc.) would provide the scope for such first articles, within the context of initial certification. In addition to traditional manufacturing competencies, for example the ability to operate within certain quality standards (e.g., ISO 9001 [40]) and environmental standards (e.g. ISO 14000 [41]) could be required. Certification could also expand into consideration of a manufacturer's ability to capture, interpret and share relevant sustainability metrics (e.g., waste generation, energy consumption, etc.). The relevant digital competencies could be promoted through module-based online training, which could also be leveraged to ensure awareness of relevant cybersecurity standards and best practices (e.g., [42–44]).

Rather than provide a disincentive for manufacturers to participate in such certification efforts, efforts will also need to be made to ensure that the value proposition of participating in the Digital Depot program is well understood by manufacturers across the supply chain. Indeed, collaboration between traditional disconnected enterprises and industries could be promoted in anticipation of future needs to 'link' these entities in order to produce novel product lines during a crisis event. To this end, strategic networking among relevant industries and companies prior to a pandemic or similar catastrophe would be highly advantageous, and could result in significantly reduced transition and response times, were a crisis event to occur.

Across supply chains, management of access to Depot data in the event of an emergency could be carried out in accordance with previously completed certificates. For instance, a consumer appliance manufacturer with proper training and certification could gain access to ventilator manufacturing data, since the relevant manufacturing process capabilities and competencies may overlap. This data would again include digital workforce training on key manufacturing considerations such as training modules on the assembly of critical components and tuning of electronic components. Design intent of relationships between manufacturing-induced characteristics (e.g., surface finish) and functional performance in service (e.g., airflow capacity and microbial resistance) would likewise be included to support associated engineering activities.

5. SUMMARY

Disruptions in each of the life cycle stages of the product and production impact the identified risk areas for sustainability, as summarized in Table 2 and discussed in Section 2. Product designs need to account for waste generation and lifecycle impacts even under extreme duress such as imposed by emergency situations. Disruptions in the use of manufacturing assets can lead to increases in waste, impacts across the lifecycle of products especially if quality in production is compromised. Emergency impacts, such as a reduced workforce, will require rapid changes in standard modes of operations.

Table 2: Risk areas for sustainability in the context of the functional areas of production. MFG: Manufacturing.

	Sustainability Risk Areas		
Life Cycle Stages	**MFG Operational Modes**	**Waste**	**Overall Life Cycle**
Product Design		x	x
Asset Use	x	x	x
Factory Operations	x	x	x
Supply Chain & Distribution	x	x	x
Waste Management	x	x	x
Workforce	x	x	

Further impacts can be imagined for each phase and many of these impacts can be reduced through the application of digital technologies for better planning and control. Tight control over factory operations through greater use of technologies such as simulation and design can help mitigate surges in waste generation or lifecycle impacts brought on by potential declines in product. New technologies for factory operations and maintenance that involve greater use of remote capabilities can also help to reduce unanticipated sustainability impacts. Similarly, the impact in changes to operational modes could be reduced through virtual workforce training on the production needs.

The interplay of the risk factors with the identified trends in manufacturing (environmentally-informed design and manufacturing on demand in an environment of disrupted resources) produce significant opportunities for future research directions.

6. CONCLUSION

In this manuscript, the concept of a Digital Depot is proposed as a potential alternative to traditional physical stockpiles of finished goods. Rather than over-producing and stockpiling manufactured products and components that may be needed during a crisis event in anticipation of such events, a Digital Depot could offer a more resilient and sustainable alternative in the future. The incorporation of flexible and digital manufacturing technologies would support the development of the Digital Depot by facilitating access to data and enabling agile production capabilities. By virtue of being a digitally-enabled system, the Digital Depot would require significant paradigm shifts across industry in order to deliver the desired benefits. While this present work provides and technical description of the various stages of production that would be impacted by the Depot, this framework still contains significant gaps. To further develop the proposed concept, we envision convergent future work across a variety of disciplines, ranging from economics to establish clear business cases, to social and environmental studies to further define the scope and impact of environmental and societal metrics. Overall, successful transition to a more digitally-enabled future state of 21st century manufacturing will require close collaboration between a variety of stakeholders from academia, industry, government and non-profit groups. In this context, the Digital Depot could serve as a catalyst and enabler to bring together these diverse groups to realize resilient and sustainable value creation.

REFERENCES

[1] Thomasnet.com, 2020, "Thomas COVID-19 Resource Hub" [Online]. Available: https://help.thomasnet.com/covid-19-resource-hub/. [Accessed: 18-Nov-2020].

[2] Diaz-Elsayed, N., Morris, K. C., and Schoop, J., 2020, "Realizing Environmentally Conscious Manufacturing in the Post–COVID-19 Era," Smart Sustain. Manuf. Syst., **4**(3), p. 20200052.

[3] World Health Organization, 2020, "Shortage of Personal Protective Equipment Endangering Health Workers Worldwide" [Online]. Available: https://www.who.int/news/item/03-03-2020-shortage-of-personal-protective-equipment-endangering-health-workers-worldwide. [Accessed: 18-Nov-2020].

[4] Biden, J. R., 2021, "Executive Order on America's Supply Chains" [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2021/02/24/executive-order-on-americas-supply-chains/. [Accessed: 25-Feb-2021].

[5] Friedersdorf, C., 2020, "The Government Is Failing by Doing Too Little, and Too Much," Atl. [Online]. Available: https://www.theatlantic.com/ideas/archive/2020/03/two-kinds-pandemic-failures/608767/. [Accessed: 19-Nov-2020].

[6] Agrawal, M., Dutta, S., Kelly, R., and Millan, I., 2021, "COVID-19: An Inflection Point for Industry 4.0" [Online]. Available: https://www.mckinsey.com/business-functions/operations/our-insights/covid-19-an-inflection-point-for-industry-40. [Accessed: 24-Feb-2021].

[7] Dey, M., Frazis, H., Loewenstein, M. A., and Sun, H., 2020, "Ability to Work from Home: Evidence from Two Surveys and Implications for the Labor Market in the COVID-19 Pandemic," Mon. Labor Rev. [Online]. Available: https://www.bls.gov/opub/mlr/2020/article/ability-to-work-from-home.htm. [Accessed: 19-Nov-2020].

[8] Lisa Allison, A., Ambrose-Dempster, E., Domenech Aparsi, T., Bawn, M., and Casas, M., 2020, "The Environmental Dangers of Employing Single-Use Face Masks as Part of a COVID-19 Exit Strategy," UCL Open Environ. Prepr.

[9] Fan, Y. Van, Jiang, P., Hemzal, M., and Klemeš, J. J., 2021, "An Update of COVID-19 Influence on Waste Management," Sci. Total Environ., **754**, p. 142014.

[10] Patrício Silva, A. L., Prata, J. C., Walker, T. R., Duarte, A. C., Ouyang, W., Barcelò, D., and Rocha-Santos, T., 2021, "Increased Plastic Pollution Due to COVID-19 Pandemic: Challenges and Recommendations," Chem. Eng. J., **405**, p. 126683.

[11] Bradshaw, K., 2020, "COVID-19 Pandemic Reveals True Importance of Recycling and the Supply Chain," Waste360.

[12] Wei, G., and Manyu, L., 2020, "The Hidden Risks of Medical Waste and the COVID-19 Pandemic," Waste360.

[13] Ellison, B., and Kalaitzandonakes, M., 2020, "Food Waste and Covid-19: Impacts along the Supply Chain • Farmdoc Daily," farmdoc Dly., **10**(164).

[14] Harris, G. A., 2019, *Industry Readiness for Digital Manufacturing May Not Be As We Thought*.

[15] Harris, G., Yarbrough, A., Abernathy, D., and Peters, C., 2019, "Manufacturing Readiness for Digital Manufacturing," Manuf. Lett., **22**, pp. 16–18.

[16] Mittal, S., Khan, M. A., Romero, D., and Wuest, T., 2018, "A Critical Review of Smart Manufacturing & Industry 4.0 Maturity Models: Implications for Small and Medium-Sized Enterprises (SMEs)," J. Manuf. Syst., **49**, pp. 194–214.

[17] Federal Emergency Management Agency, 2018, "Continuity Guidance Circular - February 2018" [Online]. Available: https://www.fema.gov/sites/default/files/ 2020-07/Continuity-Guidance-Circular_031218.pdf. [Accessed: 19-Nov-2020].

[18] El-Dardiry, A., 2020, "How to Leverage Digitization and Reduce Freight Volatility," Supply Chain Dive [Online]. Available: https://www.supplychaindive.com/news/how-to-leverage-digitization-and-reduce-freight-volatility/570956/. [Accessed: 19-Nov-2020].

[19] Vanapalli, K. R., Sharma, H. B., Ranjan, V. P., Samal, B., Bhattacharya, J., Dubey, B. K., and Goel, S., 2021, "Challenges and Strategies for Effective Plastic Waste Management during and Post COVID-19 Pandemic," Sci. Total Environ., **750**, p. 141514.

[20] Lu, Y., Morris, K. C., and Frechette, S., 2016, *Current Standards Landscape for Smart Manufacturing Systems*, National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD, USA.

[21] International Organization for Standardization, 2020, "ISO/PRF 10303-1 - Industrial Automation Systems and Integration — Product Data Representation and Exchange — Part 1: Overview and Fundamental Principles" [Online]. Available: https://www.iso.org/standard/72237.html. [Accessed: 20-Nov-2020].

[22] Letelier, K., 2019, "Time to Push for Step and Other Open Standards" [Online]. Available: https://www.sme.org/time-to-push-for-step-and-other-open-standards. [Accessed: 20-Nov-2020].

[23] International Organization for Standardization, 2020, "ISO/DIS 23247-1 - Automation Systems and Integration — Digital Twin Framework for Manufacturing — Part 1: Overview and General Principles" [Online]. Available: https://www.iso.org/standard/75066.html. [Accessed: 20-Nov-2020].

[24] International Organization for Standardization, 2016, *ISO 14040:2006 - Environmental Management -- Life Cycle Assessment -- Principles and Framework.*

[25] International Organization for Standardization, 2020, "ISO/TS 14027:2017 Environmental Labels and Declarations — Development of Product Category Rules."

[26] ASTM, 2020, "Subcommittee E60.13 : Published Standards under E60.13 Jurisdiction" [Online]. Available: https://www.astm.org/COMMIT/SUBCOMMIT/E6013.htm. [Accessed: 20-Nov-2020].

[27] Mani, M., Larborn, J., Johansson, B., Lyons, K. W., and Morris, K. C., 2016, "Standard Representations for Sustainability Characterization of Industrial Processes," J. Manuf. Sci. Eng. Trans. ASME, **138**(10).

[28] Komoto, H., Bernstein, W. Z., Kwon, S., and Kimura, F., 2020, "Standardizing Environmental Performance Evaluation of Manufacturing Systems through ISO 20140," *Procedia CIRP*, Elsevier B.V., pp. 528–533.

[29] International Organization for Standardization, 2019, *ISO 20140-1:2019*.

[30] Kellens, K., Baumers, M., Gutowski, T. G., Flanagan, W., Lifset, R., and Duflou, J. R., 2017, "Environmental Dimensions of Additive Manufacturing: Mapping Application Domains and Their Environmental Implications," J. Ind. Ecol., **21**(S1), pp. S49–S68.

[31] Bernstein, W. Z., Tensa, M., Praniewicz, M., Kwon, S., and Ramanujan, D., 2020, "An Automated Workflow for Integrating Environmental Sustainability Assessment into Parametric Part Design through Standard Reference Models," Procedia CIRP, **90**, pp. 102–108.

[32] Liebsch, T., 2020, "The Rise of the Face Mask: What's the Environmental Impact of 17 Million N95 Masks? - Ecochain," Ecochain Meas. Sustain. [Online]. Available: https://ecochain.com/knowledge/footprint-face-masks-comparison/. [Accessed: 19-Oct-2020].

[33] Klemeš, J. J., Fan, Y. Van, and Jiang, P., 2020, "The Energy and Environmental Footprints of COVID-19 Fighting Measures – PPE, Disinfection, Supply Chains," Energy, **211**, p. 118701.

[34] Zhou, L., Li, J., Li, F., Meng, Q., Li, J., and Xu, X., 2016, "Energy Consumption Model and Energy Efficiency of Machine Tools: A Comprehensive Literature Review," J. Clean. Prod., **112**, pp. 3721–3734.

[35] Diaz, N., Helu, M., Jayanathan, S., Chen, Y., Horvath, A., and Dornfeld, D., 2010, "Environmental Analysis of Milling Machine Tool Use in Various Manufacturing Environments," *Proceedings of the 2010 IEEE International Symposium on Sustainable Systems and Technology, ISSST 2010*.

[36] Diaz-Elsayed, N., Dornfeld, D., and Horvath, A., 2015, "A Comparative Analysis of the Environmental Impacts of Machine Tool Manufacturing Facilities," J. Clean. Prod., **95**, pp. 223–231.

[37] Bachinski, T. J., and Donaldson Co Inc, 1993, "Pleated Filter Media Having a Continuous Bead of Adhesive between Layers of Filtering Material."

[38] Schoop, J., Adeniji, D., and Brown, I., 2019, "Computationally Efficient, Multi-Domain Hybrid Modeling of Surface Integrity in Machining and Related Thermomechanical Finishing Processes," *Procedia CIRP*, Elsevier B.V., pp. 356–361.

[39] Astheimer, R., Re, K. Del, Gopalakrishnan, S., Hartman, N., and Sangid, M., 2019, "Extending Model Based

Morris, KC; Diaz-Elsayed, Nancy; Schoop, Julius. "Towards a Digital Depot to Support Sustainable Manufacturing During Crisis Response."
Presented at ASME Manufacturing Science and Engineering Conference 2021 (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021.

Definition to Capture Product Behavior and Contextual Information Using a Model Based Feature Information Network," CAD Solutions, LLC, pp. 253–257.

[40] International Organization for Standardization, 2020, "ISO 9000 Family Quality Management" [Online]. Available: https://www.iso.org/iso-9001-quality-management.html. [Accessed: 20-Nov-2020].

[41] International Organization for Standardization, *ISO 14000 Family – Environmental Management*.

[42] Shackelford, S. J., Proia, A., Martell, B., and Craig, A., 2015, "Toward a Global Cybersecurity Standard of Care? Exploring the Implications of the 2014 NIST Cybersecurity Framework on Shaping Reasonable National and International Cybersecurity Practices," Tex. Int. Law J., (291), pp. 1–58.

[43] Stouffer, K. A., Zimmerman, T. A., Tang, C., Lubell, J., Cichonski, J. A., and McCarthy, J., 2017, *Cybersecurity Framework Manufacturing Profile*.

[44] IT Governance USA Inc., 2020, "Cybersecurity Standards List" [Online]. Available: https://www.itgovernanceusa.com/cybersecurity-standards. [Accessed: 20-Nov-2020].

Combinatorial Testing Metrics for Machine Learning

Erin Lanus*, Laura J. Freeman*, D. Richard Kuhn†, Raghu N. Kacker†
*Hume Center for National Security and Technology, Virginia Tech, Arlington, VA, USA {lanus, laura.freeman}@vt.edu
†National Institute of Standards and Technology, Gaithersburg, MD, USA {kuhn, raghu.kacker}@nist.gov

***Abstract*—This paper defines a set difference metric for comparing machine learning (ML) datasets and proposes the difference between datasets be a function of combinatorial coverage. We illustrate its utility for evaluating and predicting performance of ML models. Identifying and measuring differences between datasets is of significant value for ML problems, where the accuracy of the model is heavily dependent on the degree to which training data are sufficiently representative of data encountered in application. The method is illustrated for transfer learning without retraining, the problem of predicting performance of a model trained on one dataset and applied to another.**

***Index Terms*—combinatorial testing, machine learning, operating envelopes, transfer learning, test set selection**

I. Introduction

In software and hardware, component systems are often well designed and tested, but failures occur during integration due to unexpected interactions between components. A study [1] of empirical data found that nearly all failures in software are caused by a limited number of interacting components and concluded that testing interactions of between four and six components could detect all failures in the software systems considered. This result has led to broader adoption of combinatorial testing, because it showed that strong assurance could be achieved without exhaustive testing of software and hardware systems [2].

Conducting testing of systems with embedded machine learning (ML) using conventional software approaches poses challenges due to characteristics of ML such as the large input space, effort required for white box testing, and emergent behaviors apparent only at integration or system levels [3], [4]. CT is a black box approach to testing an integrated system using a pseudo-exhaustive strategy for large input spaces. Thus far, CT has been applied to test case generation for autonomous vehicle systems with embedded ML components [5], testing the internal state space of a neural network [6], feature selection [7], and explainable ML [8].

An ML model is trained on *examples* consisting of *values* assigned to *features* and possibly a *label*, such as the membership class for the example. The data is fundamental to ML model performance. In this paper, we leverage CT for testing ML systems through comparison of datasets to consider how differences between members of two classes lead to classification decisions and if differences between datasets are useful for predicting whether a model trained on one dataset will perform as expected on another. Comparing datasets via combinations of features is possible at three levels of granularity: 1) the count of combinations that are present or absent, 2) which specific combinations are present or absent, and 3) the distribution of combinations.

In this work, we define a new combinatorial coverage metric for comparing ML datasets in § II focusing on the first level of granularity (presence/absence of combinations). We highlight two distinct areas of applications of the metric in § III. The metric's utility based on interpretable features in the data is demonstrated for fault localization and explainable classification. The use of the metric to define a model's operating envelope extends to applications in transfer learning, selection of training and test datasets, and directing data collection and labeling efforts. We discuss problems for future work in § IV.

II. Metrics

We treat features, including the label when available, as *factors* for CT. Continuous-valued factors must be discretized prior to applying CT so that each factor has a corresponding finite set of values. A t-way *value combination* is an assignment of specific values to t of the factors, or a t-tuple of (factor, value) pairs. If there are k factors, each example then contains $\binom{k}{t}$ factor combinations with one value combination each.

Combinatorial coverage, also called total t-way coverage, is a metric from the CT literature [9] to describe the proportion of valid t-way value combinations appearing in a set (Fig. 1). Value combinations that appear in the set are *covered* by the set. Define a universe with k factors and their respective values so that $\mathcal{U}$ is the set of all valid examples, and let $\mathcal{U}_t$ be the set of valid t-way value combinations. If some value combination is invalid, it is a *constraint* and can be removed from $\mathcal{U}_t$. Given a dataset $\mathcal{D} \subseteq \mathcal{U}$, define $\mathcal{D}_t$ as the set of t-way value combinations appearing in $\mathcal{D}$. (We acknowledge a slight abuse of notation as $\mathcal{D}$ may be a multiset. This does not impact the metrics.) Denote set cardinality by $|\mathcal{D}_t|$. The t-way combinatorial coverage [9] of $\mathcal{D}$ is

$$CC_t(\mathcal{D}) = \frac{|\mathcal{D}_t|}{|\mathcal{U}_t|}.$$

Let $\mathcal{S}$ and $\mathcal{T}$ be datasets and define $\mathcal{S}_t, \mathcal{T}_t$ as the set of t-way value combinations appearing in $\mathcal{S}, \mathcal{T}$, respectively. The set difference $\mathcal{T}_t \backslash \mathcal{S}_t$ is the set of value combinations appearing

Research of EL and LJF funded in part by MITRE University Innovation Exchange program.

Lanus, Erin; Freeman, Laura; Kuhn, D. Richard; Kacker, Raghu N. "Combinatorial Testing Metrics for Machine Learning." Presented at 14th IEEE International Conference on Software Testing, Verification and Validation (ICST Workshops 2021). April 12, 2021 - April 16, 2021.

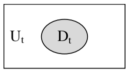


Fig. 1. Venn diagram showing $CC_t(\mathcal{D})$ as the coverage of $\mathcal{U}_t$ by $\mathcal{D}_t$.

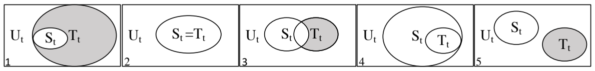


Fig. 2. Venn diagrams of the set theoretic relationships between $\mathcal{S}$ and $\mathcal{T}$.

in $\mathcal{T}_t$ but not in $\mathcal{S}_t$. We define the t-way set difference combinatorial coverage

$$SDCC_t(\mathcal{T} \setminus \mathcal{S}) = \frac{|\mathcal{T}_t \setminus \mathcal{S}_t|}{|\mathcal{T}_t|}$$

as the proportion of t-way value combinations appearing in $\mathcal{T}$ but not $\mathcal{S}$. Constraints need not be explicitly defined as only value combinations present in $\mathcal{T}$ are considered. $SDCC_t$ is a score between 0 and 1 inclusive. The set theoretic relationships (Fig. 2) and corresponding ranges of $SDCC_t$ are:

1) $\mathcal{S}_t \subset \mathcal{T}_t \implies 0 < SDCC_t(\mathcal{T} \setminus \mathcal{S}) < 1$,
2) $\mathcal{S}_t = \mathcal{T}_t \implies SDCC_t(\mathcal{T} \setminus \mathcal{S}) = 0$,
3) $(\mathcal{S}_t \not\subset \mathcal{T}_t) \wedge (\mathcal{T}_t \not\subset \mathcal{S}_t) \wedge (\mathcal{S}_t \bigcap \mathcal{T}_t \neq \emptyset) \implies$ $0 < SDCC_t(\mathcal{T} \setminus \mathcal{S}) < 1$,
4) $\mathcal{T}_t \subset \mathcal{S}_t \implies SDCC_t(\mathcal{T} \setminus \mathcal{S}) = 0$,
5) $\mathcal{S}_t \bigcap \mathcal{T}_t = \emptyset \implies SDCC_t(\mathcal{T} \setminus \mathcal{S}) = 1$.

Set difference combinatorial coverage is directed; $SDCC_t(\mathcal{T} \setminus \mathcal{S})$ may not be equal to $SDCC_t(\mathcal{S} \setminus \mathcal{T})$. As a difference metric, higher values correspond to a larger difference between the first and second sets.

At the coarsest level of granularity, coverage is represented as a single score or Venn diagram. To provide more information, the value combinations not appearing in $\mathcal{D}_t$ for CC_t and value combinations in the set difference $\mathcal{T}_t \setminus \mathcal{S}_t$ for $SDCC_t$ are listed or plotted as status per value combination. A heatmap of value combination frequency for CC_t and difference in relative frequency for $SDCC_t$ provides the finest granularity.

III. Applications

A. Fault localization

Set differencing of t-way value combinations has been applied to the problem of fault localization. A variety of set theoretic operations can be used in reducing the set of possible failure-triggering value combinations in deterministic software [10]. Running a test set typically results in a large number of passing tests and a small number of failing tests, but only a small subset of value combinations in the failing tests will induce a failure. For P_t = value combinations in passing tests and F_t = value combinations in failing tests and C_t = fault-triggering value combinations, the first step in identifying failure-triggering value combinations is a basic elimination rule: compute $F_t \setminus P_t$, value combinations in failing tests that are not in any passing tests, which for deterministic systems must contain the fault-triggering value combinations C_t. Basic set operations can also be used to further reduce the possible value combinations involved in a failure. For example, a value combination continuity rule says that if a particular t-way value combination in F_t is included in all higher strength value combinations that contain the same t factors, then the t-way value combination is sufficient to detect the error.

B. Explainable Classification

From a certain perspective, the problem of classification in ML is essentially the same as the fault localization problem in CT. We seek to identify a small subset of factors that distinguish the class from examples not in the class. This process could be viewed as generalizing the fault localization problem, where the failing tests are the class and passing tests are non-class members – what value combinations of factor values are unique to the failing tests?

This simple observation leads to a method of producing explanations or justifications of ML classifications [11] to achieve explainable AI (XAI), by computing $C_t \setminus N_t$, the set of t-way value combinations that appear in members of the class C which are not in the non-class members of N, or are more strongly associated with C than N. For example, applying this method in a database of animal characteristics produces seven predicates that are unique to reptiles (within this database): *not aquatic AND not toothed AND four legs*, *egg-laying AND not aquatic AND four legs*, etc. These value combinations have an obvious mapping with simple rules: "if non aquatic AND not toothed AND ... ". No single-factor or 2-way value combinations are uniquely associated with the reptile class, but including 3-way value combinations makes it possible to identify class members.

Previous model induction methods have been developed to reverse engineer an explanation or model from ML output [12], [13], using statistical methods to identify characteristics most closely associated with a class. The combinatorial XAI method extends this approach by producing combinations of characteristics for explanation. This distinction is important because closely associated single factors are not necessarily contained in identifying value combinations. Rule-based expert systems are often considered easy to explain but generally are not as proficient as more opaque methods such as neural networks [14]. The combinatorial approach to XAI provides a natural mapping to clearly understandable diagnostic rules.

C. Model Operating Envelope

Computer vision includes tasks such as detecting or classifying an *object* in an image. The complexity of the domain – all of the variables affecting the production of an image – leads to high likelihood of interaction effects. Consider the problem of detecting a white truck in an image. A white truck against a light background at noon from an overhead view likely presents a more difficult detection scenario than a white truck against the same light background in late afternoon where shadows are present or from profile such that the horizon line breaks up the background. The operating envelope of an ML model describes the *contexts* in which it is expected to perform correctly; deploying to contexts outside

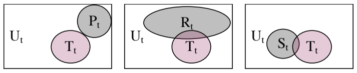


Fig. 3. Set differences used to select a source for a target from a model zoo.

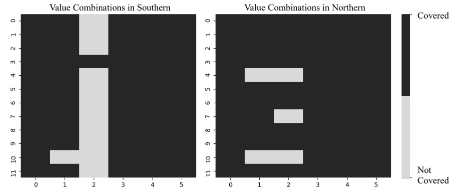


Fig. 4. Coverage of 2-way label-centric value combinations indexed by combination (y-axis) and value combination within a combination (x-axis).

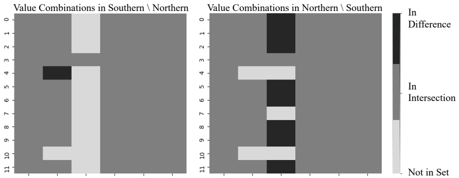


Fig. 5. Set differences of 2-way label-centric value combinations shaded by set membership; $\mathcal{T} \setminus \mathcal{S}$ is dark, $\mathcal{T} \bigcap \mathcal{S}$ is medium, and $\neg T$ is light.

of the envelope can lead to unexpected outcomes. An ML model learns about examples on which it trains, so to perform as expected in each of these contexts, it is anticipated that “enough” representative examples must be included in the training dataset. The challenge is how to define contexts and measure representativeness of the training examples.

One dimension of the operating envelope of a computer vision model is defined by describing the contexts in which the model trained as coverage of value combinations among features present in the dataset. These features may be derived directly from the image data, but there are two benefits of using metadata such as “Time of Day” or “Location” collected along with the image acting as a surrogate for contexts present in the image. Metadata are more understandable by human operators; “Time of Day” as a surrogate for lighting effects in the image is more interpretable than presenting values for luminance and contrast. Metadata may be available when image data is not, such as the case when an event is occurring in the near future in a new deployment environment for which no images have been collected. Expected factors such as “Time of Day” and “Location” can be extracted from the event profile.

When class labels are available, we describe a special way of calculating value combinations. *Label centrism* forces all value combinations to include a label; a label-centric value combination includes the label and $t-1$ of the other features. Label centrism describes the contexts in which objects appear.

Claims of representativeness by a training dataset often rely on randomized selection or counts by object type, but may fail to be representative of larger contexts in the deployment environment. Combinatorial coverage (CC) computed on a training dataset provides a measurement of the contexts on which the model trained via value combinations given the tunable parameter t. In the case of transfer learning, a model trained in one environment is deployed to a new environment, possibly without retraining or fine tuning. Where CC is a measure of coverage by a dataset with respect to some defined universe, the new metric, $SDCC$, describes a directed difference between two datasets and is useful for measuring the distance between a source dataset $\mathcal{S}$ where the model is trained and a target dataset $\mathcal{T}$ where the model is deployed. When multiple source models are available in a *model zoo*, the source dataset $\mathcal{S}$ with the smallest $SDCC_t(\mathcal{T} \setminus \mathcal{S})$ provides the best coverage of contexts in the target by the source (Fig. 3). Additionally, as value combinations in a set difference describe contexts unseen by the trained model, the list of value combinations in the set difference provides a mechanism for directing data collection or labeling efforts to include examples containing these value combinations.

A use case for the set difference application to operating envelopes for transfer learning is demonstrated on the “Planes in Satellite Imagery” Kaggle dataset [15]. The dataset is intended for binary classification and is comprised of images that either have a plane or do not have a plane along with metadata indicating the location as Northern California or Southern California. If a model is trained on the Southern subset of data $\mathcal{S}$, a performance drop occurs when used to make predictions on the Northern subset of data $\mathcal{T}$, indicating a transfer learning problem. The drop is not noted when the direction of transfer is reversed. We apply our metrics to highlight differences between the datasets that might be responsible. Twelve features are derived from the image data (the mean and variance each for the red, green, blue, hue, saturation, and luminance) and values for each feature are discretized by forming three bins encompassing equal-sized ranges. Value combinations are label-centric and $t = 2$. The Southern set contains 21,151 images and the Northern set contains 10,849 images. The $CC_2(\mathcal{S}) = \frac{60}{72} = 0.83$ and $CC_2(\mathcal{T}) = \frac{67}{72} = 0.93$, meaning that the Northern set covers more of the universe than the Southern set despite having half as many images. Fig. 4 plots the coverage of value combinations in the sets side by side.

The utility of CC for comparing a source and target pair is limited. Suppose $\mathcal{S}'_t$ contains all value combinations in the left half of a given plot and none in the right half, while $\mathcal{T}'_t = \mathcal{U}_t \setminus \mathcal{S}'_t$ contains the complement. Both have CC_2 values of 0.5. Suppose $\mathcal{S}'' = \mathcal{T}''$ yet $CC_2(\mathcal{S}'') = 0.25$. The relationship between the respective sets is not apparent via CC, which is the limitation for which $SDCC$ is designed. For the Planesnet datasets, $SDCC_2(\mathcal{S} \setminus \mathcal{T}) = \frac{1}{60} = 0.02$ and $SDCC_2(\mathcal{T} \setminus \mathcal{S}) = \frac{8}{67} = 0.12$ (Fig. 5). For this dataset, $SDCC_2$ is correlated with a drop in performance in transfer learning without retraining.

Combinatorial coverage is useful in testing for deterministic failures in software systems where the appearance of a value combination among components in one test is sufficient to cause a failure; if the components will interact to cause a

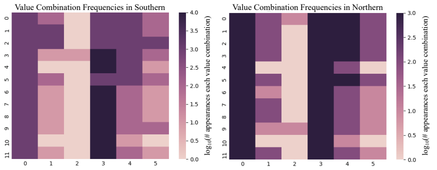


Fig. 6. Frequencies of value combinations provide distributional information.

failure, this is detected by a test suite containing that value combination at least once. Statistical learning does not have this property. We propose the CC and $SDCC$ metrics as tools to identify contexts in the target environment that are not likely to be within the model's operating envelope. However, as models are trained by updating weights each time these contexts are seen, we suspect that distribution of coverage would improve the operating envelope description. Frequently appearing value combinations indicate contexts on which the model was well trained; they could also indicate instances of overfitting. Infrequently appearing value combinations indicate contexts on which the model trained less; they could present contexts in which the model has difficulty making classifications. Our work measures and plots this distribution (Fig. 6).

D. Test Set Design

Datasets are partitioned into training $\mathcal{S}$, validation, and testing $\mathcal{T}$ sets. When datasets are large and random selection is applied, the hope is that the test set is representative of the training set as they are drawn from the same population. Computing $SDCC_t(\mathcal{S} \setminus \mathcal{T})$ and $SDCC_t(\mathcal{T} \setminus \mathcal{S})$ provides assurance against a bad random draw. A simple randomized algorithm makes several random partitions and keeps the one with the lowest $SDCC$ values. This is equivalent to testing within the operating envelope of the model.

Another testing strategy is to identify where the model fails to generalize to new contexts it has not trained by selecting test sets outside of the envelope. In this case, selecting $\mathcal{T}$ so that $SDCC_t(\mathcal{T} \setminus \mathcal{S})$ is close to 1 creates a test set containing many untrained contexts. The importance of the reverse direction for this strategy is not as clear. When $SDCC_t(\mathcal{T} \setminus \mathcal{S}) = 1$, the sets T_t and S_t are disjoint and $SDCC_t(\mathcal{S} \setminus \mathcal{T}) = 1$ necessarily. When $SDCC_t(\mathcal{S} \setminus \mathcal{T}) < 1$, the score depends on $|S_t|$.

IV. Conclusions and Future Work

This work discusses metrics that provide tools for explaining classification outcomes and defining the domain over which an ML model is expected to operate successfully. Future work is needed to test the hypothesis that models trained on source sets with smaller $SDCC_t$ distances to the target perform better in the target environment, as well as explore the usefulness of these metrics across multiple ML domains, the impact of label centrism, and choosing a "good" value combination size t.

Additionally, the sensitivity of these metrics to feature or metadata selection is critical. In the classification application, the features were directly explainable. In the computer vision application, the research had to first hypothesize reasonable features. The process of hypothesizing features, conducting initial screening experiments to select the meaningful features, and confirming results should be codified to ensure that this work is not subject to confirmation biases of the research team or over interpretation of correlations as explanatory variables.

Finally, additional work is needed to exploit the deeper levels of explainability, that is, which specific value combinations are present or absent and the distribution of those value combinations. The specific value combinations present or absent should be explored for potential explanation of how and why models perform well or poorly, potential biases introduced into the models, and predictive capabilities to new operating envelopes. Set difference frequency metrics should be developed and their application to transferability evaluated.

Acknowledgment: We thank Stephen Adams, Sayyed Ahamed, Peter Beling, Tyler Cody, Adam Edwards, and Sachin Shetty for the motivating ML work on the Planesnet dataset. We thank the reviewers for their helpful comments.

Disclaimer: Any mention of product names does not imply endorsement by NIST nor that the products mentioned are necessarily the best available for the purpose.

References

[1] D. R. Kuhn, D. R. Wallace, and A. M. Gallo, "Software fault interactions and implications for software testing," *IEEE Transactions on Software Engineering*, vol. 30, no. 6, pp. 418–421, 2004.

[2] C. Nie and H. Leung, "A survey of combinatorial testing," *ACM Computing Surveys (CSUR)*, vol. 43, no. 2, pp. 1–29, 2011.

[3] D. Marijan, A. Gotlieb, and M. Kumar Ahuja, "Challenges of testing machine learning based systems," in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 2019, pp. 101–102.

[4] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, pp. 1–36, 2020.

[5] C. E. Tuncali, G. Fainekos, H. Ito, and J. Kapinski, "Simulation-based adversarial test generation for autonomous vehicles with machine learning components," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1555–1562.

[6] L. Ma, F. Juefei-Xu, M. Xue, B. Li, L. Li, Y. Liu, and J. Zhao, "Deepct: Tomographic combinatorial testing for deep learning systems," in *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2019, pp. 614–618.

[7] S. Vilkomir, J. Wang, N. L. Thai, and J. Ding, "Combinatorial methods of feature selection for cell image classification," in *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2017, pp. 55–60.

[8] R. Kuhn and R. Kacker, "An application of combinatorial methods for explainability in artificial intelligence and machine learning (draft)," National Institute of Standards and Technology, Tech. Rep., 2019.

[9] D. R. Kuhn, I. D. Mendoza, R. N. Kacker, and Y. Lei, "Combinatorial coverage measurement concepts and applications," in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation Workshops*, 2013, pp. 352–361.

[10] D. R. Kuhn, R. N. Kacker, and Y. Lei, "Practical combinatorial testing," *NIST special Publication*, vol. 800, no. 142, p. 142, 2010.

[11] D. R. Kuhn, R. N. Kacker, Y. Lei, and D. E. Simos, "Combinatorial methods for explainable AI."

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

[13] F. Shakerin and G. Gupta, "Induction of non-monotonic logic programs to explain boosted tree models using lime," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3052–3059.

[14] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, no. 2, 2017.

[15] Rhammell, "Planes in satellite imagery," Jan 2018. [Online]. Available: https://www.kaggle.com/rhammell/planesnet

Lanus, Erin; Freeman, Laura; Kuhn, D. Richard; Kacker, Raghu N. "Combinatorial Testing Metrics for Machine Learning." Presented at 14th IEEE International Conference on Software Testing, Verification and Validation (ICST Workshops 2021). April 12, 2021 - April 16, 2021.

Combinatorially XSSing Web Application Firewalls

Bernhard Garn
SBA Research
Vienna, Austria
bgarn@sba-research.org

Daniel Sebastian Lang
Vienna University of Technology
Vienna, Austria
e1005115@student.tuwien.ac.at

Manuel Leithner
SBA Research
Vienna, Austria
mleithner@sba-research.org

D. Richard Kuhn
NIST
Gaithersburg, MD, USA
kuhn@nist.gov

Raghu Kacker
NIST
Gaithersburg, MD, USA
raghu.kacker@nist.gov

Dimitris E. Simos
SBA Research
Vienna, Austria
dsimos@sba-research.org

***Abstract*—Cross-Site scripting (XSS) is a common class of vulnerabilities in the domain of web applications. As it remains prevalent despite continued efforts by practitioners and researchers, site operators often seek to protect their assets using web application firewalls (WAFs). These systems employ filtering mechanisms to intercept and reject requests that may be suitable to exploit XSS flaws and related vulnerabilities such as SQL injections. However, they generally do not offer complete protection and can often be bypassed using specifically crafted exploits. In this work, we evaluate the effectiveness of WAFs to detect XSS exploits. We develop an attack grammar and use a combinatorial testing approach to generate attack vectors. We compare our vectors with conventional counterparts and their ability to bypass different WAFs. Our results show that the vectors generated with combinatorial testing perform equal or better in almost all cases. They further confirm that most of the rule sets evaluated in this work can be bypassed by at least one of these crafted inputs.**

***Index Terms*—combinatorial testing, security testing, web application, xss, web application firewall**

I. Introduction

Cross-Site Scripting (XSS) is a form of injection attack where an adversary supplies malicious input to a web application that is later transmitted to and executed in the context of the victim's browser (commonly as JavaScript code). Typically, this type of attack is possible when an application requires user input, but does not validate or sanitize this data to make sure that only safe input is processed.

Despite sustained attention from both academic and industrial researchers and consistent visibility amongst developers through resources such as the Open Web Application Security Project's (OWASP) list of *Top 10 Web Application Security Risks* [48], XSS flaws remain a widely prevalent and critical security issue for web applications.

To protect sites against exploitation of these vulnerabilities, operators often rely on web application firewalls (WAFs) [20]. These security systems operate on HTTP traffic and filter malicious requests before they can reach the actual application. However, studies on the effectiveness of WAFs show that they do not provide perfect protection from SQL injection attacks [1]. It is not yet clear to what degree a WAF can prevent XSS flaws from being exploited.

A large body of research deals with different aspects of XSS attacks and defensive mechanisms [20]. A variety of software testing methods have been used in order to improve XSS vulnerability detection, including search-based testing [5], unit testing [26], mutation-based testing [24], [39] as well as evolutionary approaches [12], [13], [45].

Web vulnerability scanners provide an automated way to identify XSS flaws and other security issues. However, research has shown that these products have low detection capabilities and a high rate of false positives (in other words, they often classify unsuccessful attacks as successful). Moreover, these tools cannot guarantee a certain coverage of the input space, making it difficult to estimate the reliability of the results [17]. While many works in the literature were successful in exploiting XSS vulnerabilities in web applications, few consider the presence of additional security mechanisms such as WAFs.

In this paper, we propose an attack model for the combinatorial derivation of XSS attack vectors with the goal of *bypassing* or *evading* the filters imposed by WAFs. From a software testing perspective, we consider a WAF as the *system under test* (SUT). We do not develop specific exploits against the web applications protected by these systems. In a case study, we evaluate the performance of the XSS attack vectors generated using the method described in this work. We further compare them against three traditional state-of-the-art lists of XSS attack vectors. Our results show that our approach performs as well or better against almost all SUTs considered herein.

Combinatorially instantiated attack grammars have previously been used for the creation of XSS attack vectors targeting web applications [8], [9], [19], [41]. In contrast, the

SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.

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.

Garn, Bernhard; Lang, Daniel; Leithner, Manuel; Kuhn, D. Richard; Kacker, Raghu N.; Simos, Dimitris. "Combinatorially XSSing Web Application Firewalls." Presented at 14th IEEE International Conference on Software Testing, Verification and Validation (ICST Workshops 2021). April 12, 2021 - April 16, 2021.

method presented in this work evaluates the ability of WAFs to correctly classify malicious requests, thus exploring a novel application domain of combinatorial security testing [40].

Contribution. In particular, this paper makes the following contributions:

- An attack grammar modelling XSS attack vectors for bypassing WAFs,
- Evaluation of combinatorial test sets of different strengths based upon the proposed grammar in a case study against seven SUTs,
- Performance comparison between results achieved by combinatorial test sets of XSS attack vectors and traditional static state-of-the-art lists of attack vectors.

This paper is structured as follows. In Section II, we introduce basic concepts used throughout this work. Section III explains the process and components of our proposed testing approach as well as the developed attack model. In Section IV, we introduce the chosen SUTs, our experimental setup, as well as the selection of static lists of attack vectors for comparison. The following Section V presents and evaluates the results of the case study. Section VI gives an overview of related work, while Section VII contains a discussion of potential threats to validity. Finally, Section VIII concludes this work and provides an outlook for future research.

II. Preliminaries

This section gives an overview of the basic concepts used in this work. In Section II-A, we describe web application firewalls and how they are used to protect web applications from attacks. Finally, we outline *penetration testing* and *combinatorial security testing* for XSS vulnerabilities in Section II-B.

A. Web Application Firewalls

Web application firewalls are conceptually similar to traditional firewalls, offering capabilities to filter traffic based on user-defined rules. However, instead of inspecting network traffic on OSI layer 4 and below [11], they are primarily focused on application layer traffic, particularly HTTP requests. They are often implemented as web server modules or dedicated appliances that are logically placed in front of a web server. When a WAF detects malicious input, it blocks the request and will not forward it to the application. This provides an additional layer of protection by preventing the exploitation of both known and unknown issues in vulnerable software.

When verifying the correctness of filters, it is often important to consider the complexity of the rules employed to make a decision on whether to forward, drop or otherwise process some unit of traffic. Most rules used by traditional firewalls, particularly those operating on lower layers, inspect fields with bounded values: There is a finite number of IP addresses, network protocols, ports and so on. Generating test cases for these rules is relatively simple and could potentially even be performed exhaustively (i.e. by iterating over all possible values of all inspected fields). In contrast, WAFs operate on almost entirely unconstrained user input (i.e. strings), which yields a much larger search space that would have to be considered in order to identify all possible attacks. Furthermore, deciding whether a given user input is actually malicious or not may require additional information. The string `<script>alert(1);</script>` is a classic example for an XSS attack. However, in a message board about programming or security topics, this string might also in fact be a valid content for a comment that *describes* XSS. Without additional specifications, a WAF will have to make certain assumptions about the semantics of the software it is meant to protect. A stricter configuration generally leads to a greater amount of requests that will be blocked, at the risk of increasing the chances of legitimate requests being rejected. Therefore, a balance has to be found between maximum security and little to no false positives (i.e. benign requests that were blocked in error).

A request addressed to an application first reaches the WAF. If it is deemed benign, it is forwarded to the application and processed. The corresponding response is then returned and passed through to the user. This process can be seen in Figure 1. In contrast, Figure 2 visualizes the flow of network traffic that is deemed malicious. Instead of being forwarded to the application, the WAF immediately returns a response to the user, commonly indicating that the submitted request has been denied. In this case, the (possibly malicious) request never reaches the intended target and is thus incapable of achieving exploitation.

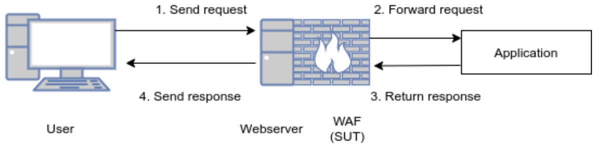


Fig. 1: Request process with benign content or undetected attack.

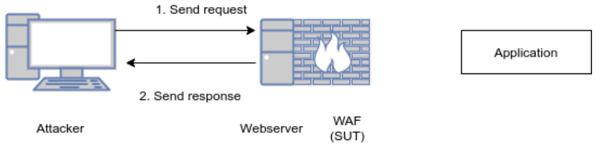


Fig. 2: Request process with detected malicious content.

B. Combinatorial Security Testing

a) Software security testing: Software security testing, an integral part of a properly implemented *software development life cycle* (SDLC) [10], [21], [25], [27], [47], [15], [33], seeks to find vulnerabilities in a given piece of software, which is usually referred to as the SUT.

b) Penetration testing: While security testing as part of an SDLC is often performed in a *white-box* setting, i.e. one where the tester has full access to the source code and internal

Garn, Bernhard; Lang, Daniel; Leithner, Manuel; Kuhn, D. Richard; Kacker, Raghu N.; Simos, Dimitris. "Combinatorially XSSing Web Application Firewalls." Presented at 14th IEEE International Conference on Software Testing, Verification and Validation (ICST Workshops 2021). April 12, 2021 - April 16, 2021.

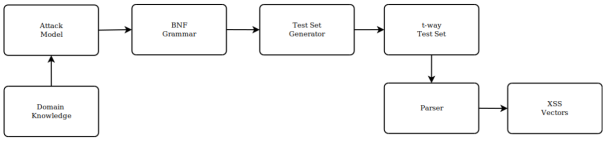


Fig. 3: Modeling process and test generation; taken from [41].

documentation of an application, *penetration testing* seeks to emulate the view of an attacker and is thus usually performed in a *black-box* setting. Additionally, penetration testing is not restricted to identifying vulnerabilities, but also encompasses an array of other activities, from the initial reconnaissance and enumeration phases to the final reporting stage. It may include a larger scope than pure software (security) testing, focusing on the overall risk to the business. [4], [34].

c) Combinatorial security testing: Combinatorial testing is a model-based approach: the SUT is described with finitely many parameters, each of which can take finitely many values. This abstract model of the SUT is termed an *input parameter model* (IPM), and parameter values for the model are commonly provided by application domain experts. When security domain knowledge is used to create an IPM modeling security properties of a system or attacks on a system to be used with CT, it is called *combinatorial security testing* (CST). This approach has been successfully applied to several different kinds of attacks in the general domain of information security [40]. The individual test cases in these generated t-way test sets constitute *abstract attack vectors*, which will subsequently be translated into a domain-specific representation to be executed against a SUT. Note that we follow the terminology used in [8] when referring to test cases as attack vectors.

To make the overall process more tangible, we illustrate the CST approach used in this work in the domain of testing for XSS vulnerabilities with an example. Figure 3 depicts the individual phases involved.

We start with a context free grammar for a test case `<test>` in BNF:

```
<test>  ::= <tag> <space> <quote>
<tag>   ::= script | img | a
<space> ::= ␣ | \n | \r \ | \0 | <empty>
<quote> ::= ′ | " | ` | <empty>
```

Strings derived from this grammar are to be interpreted as a basic and preliminary form of XSS attack vectors. Note that for testing purposes, there are many ways to sample strings in the language that this grammar generates, i.e. there are different *strategies* on how to select test cases. In this work, we employ a combinatorial sampling strategy. Specifically, this grammar can used to define an IPM with three parameters, which are given by the three non-terminal symbols `<tag>`, `<space>` and `quote`. The selected values of these parameters are then concatenated to form a single test case `<test>`. The elements of this test case may additionally undergo a *translation* step in order to transform it from an abstract test case (i.e. row in a t-way test set) to a *translated* (i.e. ready-to-be executed) test case. We call a function that performs this kind of concatenation and transformation, as well as optional mutations such as changing the case or encoding of strings, a PAYLOAD GENERATOR. In effect, the generation of XSS attack vectors consists of two steps: The first is the generation of a t-way test set based on the IPM, while the second is the construction of translated test cases by the PAYLOAD GENERATOR function. An example for a translated test case (and thus a XSS attack vector) derived from this IPM would be `script\r"`.

Note that based on the IPM specified by the grammar above, we could generate $3 \times 5 \times 4 = 60$ different test cases in total, which would correspond to an exhaustive test set. A pairwise test set, however, exists with only 20 test cases. This pairwise test set achieved a reduction of about 60% in terms of test set size compared to the full input space defined by this IPM.

III. Methodology and Attack Model

This section describes the core of our approach for testing for XSS attack vectors capable of bypassing WAFs. We present an overview of our methodology in Section III-A and discuss our developed attack model in detail in Section III-B. The description of the used testing oracle in this work in given in Section III-C, where we also explain the advantages and drawbacks of choosing the WAFs themselves as test oracles.

A. Testing process overview

Figure 3 shows an overview of the security testing methodology followed in this work, the phases of which are discussed below. Note the WAFs themselves appear twice in our testing methodology. First, they appear as SUTs, when we speak of them as *evaluation targets*. Second, they are implicitly used as oracles when their blocking decisions are used to determine whether a request was actually able to bypass them.

In order to facilitate the automated execution of our approach, we implemented a prototype Python tool that performs the test set generation and translation, execution of test cases (i.e. the submission of XSS attack vectors), and logging of results to a PostgreSQL relational database.

a) Testing environment: We use Docker[1] and Docker-compose[2] to run each SUT in an isolated container with its respective configuration. Figure 4 shows the logical network layout of the virtualized test environment. All SUTs run in their own container and are isolated from the attacking client (i.e. our prototype tool). This not only ensures that no unwanted interactions between the client and the SUTs take place, but also makes adding new SUTs very easy, since they are not tasked with additional steps such as saving evaluation results and can thus remain unmodified.

[1] https://www.docker.com

[2] https://docs.docker.com/compose/

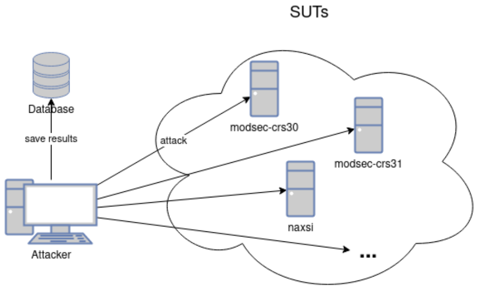


Fig. 4: Network topology.

b) Attack vector generation phase: In the attack vector generation phase, the test sets consisting of XSS attack vectors are generated. For the CST approach to XSS attack vector generation presented in this work we employ several t-way test sets of varying strength. The underlying attack grammar (i.e., IPM) will be described in Section III-B, while the created test sets are described in Section IV-C.

For the purpose of comparing the results achieved by our attack vectors with existing approaches, we also include three lists of XSS attack vectors. Details of these lists are given in Section IV-B.

c) Attack phase: Before this phase begins, the containerized SUTs that were specified in the preparation phase are launched. For one or more specified SUTs, our tool first loads the respective set of XSS attack vectors from the database and then submits them using the *requests* library for Python3 [35]. In most cases, the payload is transmitted as the value of a HTTP GET parameter, since these parameters are scanned in all rules by every tested WAF in this work. This behavior had to be adjusted in some cases; see Table II and Section IV for details. The information on whether the SUT blocks the request or passes it on is saved in our database. We provide more details on the oracle in Section III-C. Note that we ignore what happens to the request after it passes the WAF, since the decision of the oracle is already known at this point.

d) Evaluation phase: After all experiments have been performed, the results of the attack phase can be analyzed and evaluated. Details on this step are presented in Section V.

B. Attack Model

The attack grammar in the form of an IPM devised for testing WAFs for XSS vulnerabilities contains 12 parameters and is given below. The number of distinct values of the respective parameter is enclosed in parentheses.

```
FOBRACKET (1): ’<’
TAG (6): ’img’, ’script’, ’body’, ...
FCBRACKET’ (1): ’>’
QUOTE (4): ’"’, "’", ...
SPACE (9): "\n", "\t", "\r", ...
EVENT (11): ’data’, ’href’, ’src’, ...
TAG_CASE (3): ’lower’, ’upper’, ’mixed’
EVENT_CASE (3): ’lower’, ’upper’, ’mixed’
PAYLOAD (8): ’’’confirm(1)’’’,
             ’’’javascript:alert(1)’’’,
             ’’’alert‘1‘’’’, ...
LOBRACKET (1): ’</’
LCBRACKET (1): ’>’
IS_DOUBLE_ENCODED (2): "True", "False"
```

Note that this IPM contains so-called *meta-parameters* for XSS attack vector generation which, in contrast to the parameters given in the example IPM in Section II, encode specific *translation options* for the PAYLOAD GENERATOR function and do not appear in the final XSS attack vectors.

In particular, the meta-parameters in the IPM above describe and control mutations that are commonly applied to attack vectors in security testing with the goal of evading or bypassing filters employed by security software like WAFs. The meta-parameters ”TAG_CASE” and ”EVENT_CASE” control the casing for the values of the ”TAG” and ”EVENT” parameters, respectively, which can take the values of ”lower”, ”UPPER” or ”mIxEd”. Another meta-parameter, ”IS_DOUBLE_ENCODED”, determines whether to apply URL encoding during generation. As every XSS attack vector is automatically encoded at request time by the Python3 *requests* library, setting this meta-parameter to ”True” will result in the vector being URL encoded twice in total.

A simplified version of the PAYLOAD GENERATOR function used in this work is given in Algorithm 1.

Algorithm 1: A PAYLOAD GENERATOR with its execution flow controlled by a meta-parameter.

```
Function payload_generator(tag, payload, double_encoding) : string is
    item ← ”<” + tag + ”>” + payload + ”< /” + tag + ”>”;
    if double_encoding then
        item ← urlencode(item);
    end
    return item;
end
```

C. Oracle

The oracle is the component that decides if a test case constitutes a successful attack. Depending on whether this decision was correct, four different outcomes are possible:

	Oracle True	Oracle False
Actual True	True Positive	False Negative
Actual False	False Positive	True Negative

TABLE I: Classification outcomes for attack vectors.

Ideally, a WAF would produce no false negatives or false positives. In our study this means that a WAF should block all attacks (a situation that corresponds to a true positive), but pass all benign requests (thus exhibiting a true negative). Compared to similar case studies on SQL injections [1], [2], it is more difficult to determine if an XSS payload actually exploits a vulnerability, because the injected code is executed in the user's browser and not the web application itself. In some situations, this might lead to unintentional exploitation by benign vectors [44] or prevent malicious payloads from executing, e.g. due to client-side XSS protection.

To address this problem, only intentionally malicious payloads are created, effectively eliminating true negatives. Every request that is able to bypass the WAF is considered a successful attack, while a blocked request is treated as a failed attack. We selected this approach because a browser oracle – where the payload actually has to be executed in the browser – depends on browser-specific behavior and might thus lead to incorrect results. These behaviors are relevant in a real-world scenario when trying to exploit a vulnerability, but negligible for our evaluation of XSS detection capabilities of WAFs. Ideally, the WAF blocks every payload that could lead to an exploit in any browser and in any context.

Treating the WAF's decision as an absolute also has drawbacks. Since we only create intentionally malicious payloads, we know that when the oracle does not detect an attack, it must be a false negative. However, false positives become a bigger challenge. When configuring a WAF, it is desirable to minimize false positives, which might lead to legitimate requests being blocked. Given the decision mode of our oracle (which treats every blocked request as a failed attack), simply blocking every request would give a perfect score in our evaluation. We therefore have to keep in mind that our evaluation method favors restrictive WAFs. In practice, this should not be a huge concern, because false positives are a common challenge for WAFs and default rules tend to be rather lenient as a result. In Section IV-A, we explain how different WAFs decide what constitutes an attack.

IV. Case Study

In this section, we provide the remaining details of our case study, including the tested SUTs, selected static attack lists (extracted from vulnerability scanners) for comparison and computing infrastructure that was used to execute the experiments. In Section IV-A, we describe the tested WAFs and briefly comment on differences between them. Section IV-B contains static lists of attack vectors used for comparison in this work. The computing infrastructure environment used to execute experiments is described in Section IV-C.

A. Web Application Firewalls

Three open source WAFs (ModSecurity [46], NAXSI [28] and lua-rest-waf [31]) were selected as targets, some of them in multiple configurations, resulting in a total of 7 SUTs in this case study. These applications were selected due to the availability of their source code and because they implement different request classification approaches, thus ensuring diversity.

We did not consider any commercial tools, WAFs that work exclusively based on whitelisting, or candidates that employ learning-based approaches. Deploying and running tests against WAFs based on whitelisting (which block all requests by default and only allow those that have been explicitly permitted) would have required the creation of application-specific whitelists. Similarly, learning-based WAFs would have required training before performing the evaluation, a process that was considered out of scope for this work.

WAF	Web server	Rule set	method	Parameter
ModSecurity 2	Apache 2.4	CRS 3.0 XSS	GET	q
ModSecurity 2	Apache 2.4	CRS 3.1 XSS	GET	q
ModSecurity 3	Apache 2.4	CRS 3.2 XSS	GET	q
lua-resty-waf	nginx 1.17	default XSS	GET	q
NAXSI	nginx 1.17	default	GET	q
NAXSI	nginx 1.17	wordpress whitelist	POST	comment
NAXSI	nginx 1.17	drupal whitelist	POST	user_mail

TABLE II: Target SUTs for evaluation.

Table II lists the selected WAFs, together with the set of rules, request method and payload parameter used for evaluation purposes.

The mode of operation is different for each of the WAFs listed in Table II. For this reason, each WAF and its respective configuration(s) are explained below.

a) ModSecurity: ModSecurity [46] is a WAF that can be enabled as a module in Apache [43] or nginx [14] and ships with a core rule set (CRS) [30]. The rules contained therein encode whether a request is classified as malicious or not (see Section III-C). The CRS is grouped into different attack categories (XSS, SQL injection, etc.). Only XSS rules were enabled in this case study. Each rule consists of a complex regular expression to detect malicious payloads in GET/POST HTTP parameters, cookies or other headers.

The CRS has a parameter called `paranoia level` (`PL`) that assumes integer values ranging from 1 to 4, which allows the administrator to select how strict the evaluation performed by ModSecurity should be. The default level for `PL` is 1, while a level of 2, according to [29],

> [...] is advised for moderate to experienced users who desire more complete coverage, and for all installations with elevated security requirements. PL2 may cause some FPs [False Positives] which you need to handle.

Levels 3 and 4 similarly result in an increasing number of false positives. In this work, only `PL 1` was considered to avoid a high rate of false positives. This is in line with the expected deployment options in practice, as additional adjustments to rules may be necessary when using higher levels [42].

b) NAXSI: NAXSI is a WAF that works as a module on nginx webservers [28]. In contrast to ModSecurity, NAXSI does not have complex patterns to match malicious requests, instead opting to blacklist single characters, e.g., ”<”. While this results in easily comprehensible rules, it may also lead to more false positives.

c) lua-resty-waf: This WAFs [31] works as a reverse proxy on top of the OpenResty stack [51], which is an application platform built on top of nginx. Similar to ModSecurity, rules consist of regular expressions.

Strength	# test cases	Reduction (in %)
2	99	99.97
3	794	99.76
4	4,766	98.60
5	19,311	94.35
6	58,251	82.97

TABLE III: Number of test cases in generated t-way test sets and reduction vs. exhaustive.

B. *Static attack lists*

In this case study, we wish to compare the attack vectors generated with CST against those used by state-of-the-art vulnerability scanners. Unfortunately, the tight coupling between test set generation and execution makes web vulnerability scanners such as Burp Suite [32] and OWASP ZAP [52] incompatible with our oracle and prototype tool. We therefore utilize three publicly available static lists of XSS attack vectors instead, which are of comparable quality to the vectors used by the aforementioned products.

a) rsnake: Robert Hansen, a widely known security researcher who also co-authored on a book on XSS exploits [16], created a list of XSS attack vectors [36] containing 73 handcrafted items, including a wide array of evasion techniques.

b) html5sec: Html5sec[3] is a cheat sheet that points out XSS issues in connection with attributes introduced in HTML5. We utilized a list of 136 attack vectors that try to exploit every issue mentioned therein [23].

c) portswigger: The portswigger list of attack vectors [37] was created by the cyber security company of the same name, commonly known as the developer of Burp Suite. In contrast to the other two lists, the 6047 entries in this list are not manually crafted, but instead generated by combining a set of HTML tags, attributes and payloads.

C. *Experimental setup*

All experiments were performed on a PC system running Arch Linux with an AMD Ryzen CPU and 16 GB memory. The total test generation time for all $t \in \{2, 3, 4, 5, 6\}$ using *CAgen* [49] was only 30 seconds. Table III lists the sizes of the generated combinatorial test sets as well as their reduction compared to an exhaustive test set (which would require $342,144$ test cases). Considering that we achieved approximately 120 requests/second during our evaluation, executing our combinatorial test sets for all strengths resulted in a cumulative elapsed time of about 12 minutes for each SUT.

V. Evaluation

This section presents the results of the evaluation of our case study. We introduce the evaluation metrics used to interpret and analyze the results in Section V-A. In Section V-B, we analyze the results achieved by the combinatorially generated XSS attack vectors. Finally, Section V-C contains a performance comparison between these vectors and their counterparts extracted from static lists.

A. *Evaluation metrics*

We use a slightly modified definition of EXPLOITATION RATE (ER) to quantify the performance of XSS attack vectors, a metric that is well-established in the literature [8], [41]. Specifically, for a given SUT and test set $\mathcal{T}$, the ER equals the ratio of all test cases (i.e. XSS attack vectors from the test set) that have received a test oracle verdict of *true* (i.e. successfully bypassing the WAF undetected). In other words, we obtain[4]:

$$\text{ER} = \frac{\text{\# Attack vectors in } \mathcal{T} \text{ that bypass the WAF}}{\text{\# Attack vectors in test set } \mathcal{T}}$$

When comparing the ER of two different test sets against the same SUT, a *higher* value is better.

B. *Evaluation of CST results*

We are interested in evaluating the achieved ERs for the generated combinatorial test sets with varying strengths from 2 to 6 against the different SUTs in the case study. The complete evaluation results for all targets and interaction strengths are given in Table V.

An important aspect is the influence of the interaction strength on the ER. The graphs in Figure 5 show that for each SUT, the exploitation rate is nearly constant for increasing strength. Although this may seem unimpressive at first, a constant ER for test sets with an increasing number of individual XSS attack vectors (i.e., test cases) also means more successful bypasses in absolute numbers for higher-strength test sets.

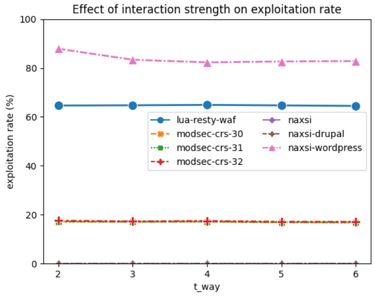


Fig. 5: Comparison of exploitation rates of combinatorial test sets for different strengths.

Two SUTs, ”NAXSI” and ”NAXSI-drupal” (NAXSI with Drupal whitelist), were able to block all attack vectors. This phenomenon arises from the unusual form of rules used by these WAFs. As explained in Section IV-A, NAXSI blacklists single characters, while all the other WAFs employ complex patterns in their rules. This leads to a high level of protection, as some characters that are common to many attack vectors (such as parenthesis, quotes or characters that start an escape sequence) are blocked. In a real-world scenario, decisions based on a single character are hardly enough to successfully detect a malicious request and will therefore yield

[3] https://www.html5sec.org

[4] They symbol # denotes the cardinality of a set.

a high number of false positives. This issue becomes very visible with the target ”NAXSI-wordpress”. The Wordpress whitelist allows several more characters in order to make sure benign requests are not blocked. However, this difference alone suffices to make this the SUT with the highest ER. In a real-world scenario, a WAF operating only on single character rules needs to work very closely with the underlying application to find a good balance between a high rate of true positives while still maintaining a low rate of false positives.

ModSecurity was the SUT with the second lowest ER for the pairwise test set of XSS attack vectors, but did not reject $\approx 17\%$ of the generated attack vectors. There was no difference in the value of the achieved ER depending on which version of the CRS had been used. Note that Table V reports a slightly lower number of total requests per combinatorial test set than what would be expected according to Table III for the SUT modsec-crs-32 (ModSecurity with CRS 3.2). This is because certain generated XSS attack vectors caused the web server to abruptly terminate the connection, leaving us unable to evaluate their effects. Based on an investigation of this issue, we suspect that this is a software bug that arises from the combination between these specific versions of the Apache web server and ModSecurity or an issue in building the relevant Docker container. However, this issue affects at most $\approx 2\%$ of test cases in each combinatorial test set. The observed behavior for this configuration of ModSecurity is otherwise similar to the results for other configurations.

C. Evaluation of static lists

Table VI shows the results for the static lists of attack vectors. Similar to our combinatorial approach, none of these vectors were able to bypass NAXSI or NAXSI-drupal, as all of them contained characters explicitly blocked by these two SUTs. As was the case for the combinatorial test sets, the number of successful injections is constant across all three ModSecurity SUTs.

Note that ERs of the portswigger list are either zero (in the case of lua-resty-waf, all three ModSecurity SUTs as well as NAXSI and NAXSI-drupal) or one (NAXSI-wordpress) in all cases. The reason for this rather unusual result is the minuscule amount of variation in this list of attack vectors.

D. Comparison between combinatorial and static test sets

Table IV lists the best ER achieved by a combinatorial test set in column `Best CT` and as well as the results achieved by the static lists. Note that the SUT modsec-crc-32 was excluded due to the observed irregularities explained above. As the results for all ModSecurity configurations are identical, they are listed as a single item.

In terms of ER, the combinatorial test sets outperform the static lists for the three SUTs lua-resty-waf, modsecurity-crc-30 and modsecurity-crc-31. For NAXSI-wordpress, the best CST result is weaker than static lists, but nevertheless still achieves a high ER of 87.9%.

Target	Best CT	rsnake	html5sec	portswigger
lua-resty-waf	64.9	6.8	7.4	0.0
ModSecurity	17.2	6.8	2.2	0.0
NAXSI-wordpress	87.9	97.3	95.6	100

TABLE IV: Exploitation rates in percent for the most important targets comparing the best combinatorial result with static lists.

target	strength	success	fail	total	ER (in %)
lua-resty-waf	2	64	35	99	64.6
modsec-crs-30	2	17	82	99	17.2
modsec-crs-31	2	17	82	99	17.2
modsec-crs-32	2	17	80	97	17.5
NAXSI	2	0	99	99	0.0
NAXSI-drupal	2	0	99	99	0.0
NAXSI-wordpress	2	87	12	99	87.9
lua-resty-waf	3	514	280	794	64.7
modsec-crs-30	3	136	658	794	17.1
modsec-crs-31	3	136	658	794	17.1
modsec-crs-32	3	136	652	788	17.3
NAXSI	3	0	794	794	0.0
NAXSI-drupal	3	0	794	794	0.0
NAXSI-wordpress	3	662	132	794	83.4
lua-resty-waf	4	3,093	1,673	4,766	64.9
modsec-crs-30	4	819	3,947	4,766	17.2
modsec-crs-31	4	819	3,947	4,766	17.2
modsec-crs-32	4	819	3,882	4,701	17.4
NAXSI	4	0	4,766	4,766	0.0
NAXSI-drupal	4	0	4,766	4,766	0.0
NAXSI-wordpress	4	3,922	844	4,766	82.3
lua-resty-waf	5	12,493	6,818	19,311	64.7
modsec-crs-30	5	3,265	16,046	19,311	16.9
modsec-crs-31	5	3,265	16,046	19,311	16.9
modsec-crs-32	5	3,265	15,800	19,065	17.1
NAXSI	5	0	19,311	19,311	0.0
NAXSI-drupal	5	0	19,311	19,311	0.0
NAXSI-wordpress	5	15,961	3,350	19,311	82.7
lua-resty-waf	6	37,565	20,686	58,251	64.5
modsec-crs-30	6	9,819	48,432	58,251	16.9
modsec-crs-31	6	9,819	48,432	58,251	16.9
modsec-crs-32	6	9,819	47,705	57,524	17.1
NAXSI	6	0	58,251	58,251	0.0
NAXSI-drupal	6	0	58,251	58,251	0.0
NAXSI-wordpress	6	48,259	9,992	58,251	82.8

TABLE V: Full evaluation results for CT suites.

VI. Related Work

Research has shown that applying XSS sanitization to input data is difficult and prone to errors [6], [38]. In [50], the authors provide an an overview of challenges to input sanitization and explore methods to this end available in web application frameworks. Even sophisticated dynamic approaches such as the method presented in [7] do not provide 100% security against such attacks. Several testing methods have been developed with the goal of bypassing such filtering mechanisms in order to improve XSS detection, including search-based testing [5], unit testing [26] and mutation-based testing [24], [39], as well as evolutionary approaches [12], [13], [45].

There has been research on SQL injection attacks with WAFs in place [1]–[3]. Estimates on the effectiveness of WAFs have been surveyed in [22].

There are several works in the literature relying on a CST approach for generating XSS attack vectors [18], establishing the applicability of CT to web application security testing. The

target	suite	success	fail	total	ER (in %)
lua-resty-waf	html5sec	10	126	136	7.4
modsec-crs-30	html5sec	3	133	136	2.2
modsec-crs-31	html5sec	3	133	136	2.2
modsec-crs-32	html5sec	3	132	135	2.2
NAXSI	html5sec	0	136	136	0.0
NAXSI-drupal	html5sec	0	272	272	0.0
NAXSI-wordpress	html5sec	130	6	136	95.6
lua-resty-waf	portswigger	0	6,047	6,047	0.0
modsec-crs-30	portswigger	0	6,047	6,047	0.0
modsec-crs-31	portswigger	0	6,047	6,047	0.0
modsec-crs-32	portswigger	0	6,047	6,047	0.0
NAXSI	portswigger	0	6,047	6,047	0.0
NAXSI-drupal	portswigger	0	12,094	12,094	0.0
NAXSI-wordpress	portswigger	6,047	0	6,047	100.0
lua-resty-waf	rsnake	5	68	73	6.8
modsec-crs-30	rsnake	5	68	73	6.8
modsec-crs-31	rsnake	5	68	73	6.8
modsec-crs-32	rsnake	5	68	73	6.8
NAXSI	rsnake	0	73	73	0.0
NAXSI-drupal	rsnake	0	146	146	0.0
NAXSI-wordpress	rsnake	71	2	73	97.3

TABLE VI: Full evaluation results for static attack vector lists.

techniques used to this end include the modelling of attack patterns found in XSS vulnerabilities [9] and the integration of constraints on the devised IPMs for a more strict generation of attack vectors [8]. Finally, an approach using locally optimized attack models is presented in [41], where the execution context was taken into account and locally optimized vectors were designed to specifically exploit certain structural properties occurring in an HTML page.

VII. Threats to Validity

In this section, we comment on possible threats to validity of this work.

First, with regards to internal validity, the primary observation is that we used a custom virtualized environment to execute our case study. This simulated network logically uses the same protocols that are also employed in an internet environment. However, we note that we have experienced unexpected behavior of one SUT (modsec-crs-32) during the execution of combinatorial test sets. Our chosen oracle does not allow to distinguish between false and true positives/negatives, and our developed grammar had to be quite conservative in order to not generate any true negatives, i.e. payloads that cannot lead to an exploit and are correctly identified as such. Since we considered any vector able to bypass a WAF as successful, this could lead to payloads skewing the results toward an overly high ER.

With regards to external validity, it is clear that the conducted case study is limited. However, we considered a diverse range of WAFs to make the resulting benchmark comparisons realistic.

VIII. Conclusion

The evaluation of the case study performed in this work shows that the exploitation rate of combinatorial test sets based on a dedicated attack grammar compares favorably to existing state-of-the-art attack sets. A practically constant ER for increasing strength of the combinatorial test sets implies an increase in the absolute number of successful attacks for increasing strength and that mere pairwise (i.e. strength $t = 2$) test sets lead to the successful bypass of some SUTs . In some cases, the combinatorial test tests outperformed static lists of XSS attack vectors considerably.

The evaluation further shows that while WAFs offer some additional protection against XSS injections to web applications, they are not infallible and may still leave web applications vulnerable to attacks. WAFs do offer some protection, even without much configuration, but they cannot substitute correct handling of input by the applications themselves.

The obtained results of the CST-approach for generating XSS attack vectors for bypassing WAFs in case study consisting of a diverse set of open source WAFs as SUTs lead to the conclusion that the CST approach has to be considered as delivering state-of-the-art performance when testing for vulnerabilities in WAFs in addition to its previous success reported in the literature in (direct) XSS attacks on web applications.

Although our results show that the CST approach proposed in this work compares favorably to traditional approaches for bypassing WAFs, there are avenues for future research. It is clear that the integration of a more sophisticated oracle would enhance the overall approach. Moreover, in future research, the CST approach proposed in this work could be refined and used to develop exploits specifically targeting a particular WAF and a single web application.

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] Dennis Appelt, Cu D. Nguyen, and Lionel Briand. Behind an Application Firewall, Are We Safe from SQL Injection Attacks? In *2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST)*, pages 1–10, Graz, Austria, April 2015. IEEE. http://ieeexplore.ieee.org/document/7102581/.

[2] Dennis Appelt, Cu D. Nguyen, Annibale Panichella, and Lionel C. Briand. A Machine-Learning-Driven Evolutionary Approach for Testing Web Application Firewalls. *IEEE Transactions on Reliability*, 67(3):733–757, September 2018. https://ieeexplore.ieee.org/document/8395015/.

[3] Dennis Appelt, Annibale Panichella, and Lionel Briand. Automatically Repairing Web Application Firewalls Based on Successful SQL Injection Attacks. In *2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE)*, pages 339–350, Toulouse, October 2017. IEEE. http://ieeexplore.ieee.org/document/8109099/.

[4] B. Arkin, S. Stender, and G. McGraw. Software penetration testing. *IEEE Security Privacy*, 3(1):84–87, 2005.

[5] Andrea Avancini and Mariano Ceccato. Security Testing of Web Applications: A Search-Based Approach for Cross-Site Scripting Vulnerabilities. In *2011 IEEE 11th International Working Conference on Source Code Analysis and Manipulation*, pages 85–94, Williamsburg, VA, USA, September 2011. IEEE. http://ieeexplore.ieee.org/document/6065200/.

[6] Davide Balzarotti, Marco Cova, Vika Felmetsger, Nenad Jovanovic, Engin Kirda, Christopher Kruegel, and Giovanni Vigna. Saner: Composing Static and Dynamic Analysis to Validate Sanitization in Web Applications. In *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, pages 387–401, Oakland, CA, USA, May 2008. IEEE. http://ieeexplore.ieee.org/document/4531166/.

[7] Prithvi Bisht and V. N. Venkatakrishnan. XSS-GUARD: Precise Dynamic Prevention of Cross-Site Scripting Attacks. In Diego Zamboni, editor, *Detection of Intrusions and Malware, and Vulnerability Assessment*, volume 5137, pages 23–43. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. http://link.springer.com/10.1007/978-3-540-70542-0_2.

[8] Josip Bozic, Bernhard Garn, Ioannis Kapsalis, Dimitris Simos, Severin Winkler, and Franz Wotawa. Attack Pattern-Based Combinatorial Testing with Constraints for Web Security Testing. In *2015 IEEE International Conference on Software Quality, Reliability and Security*, pages 207–212, Vancouver, BC, Canada, August 2015. IEEE. http://ieeexplore.ieee.org/document/7272934/.

[9] Josip Bozic, Dimitris E. Simos, and Franz Wotawa. Attack pattern-based combinatorial testing. In *Proceedings of the 9th International Workshop on Automation of Software Test - AST 2014*, pages 1–7, Hyderabad, India, 2014. ACM Press. http://dl.acm.org/citation.cfm?doid=2593501.2593502.

[10] B. Chess and B. Arkin. Software security in practice. *IEEE Security Privacy*, 9(2):89–92, 2011.

[11] John D Day and Hubert Zimmermann. The osi reference model. *Proceedings of the IEEE*, 71(12):1334–1340, 1983.

[12] Fabien Duchene, Roland Groz, Sanjay Rawat, and Jean-Luc Richier. XSS Vulnerability Detection Using Model Inference Assisted Evolutionary Fuzzing. In *2012 IEEE Fifth International Conference on Software Testing, Verification and Validation*, pages 815–817, Montreal, QC, Canada, April 2012. IEEE. http://ieeexplore.ieee.org/document/6200193/.

[13] Fabien Duchene, Sanjay Rawat, Jean-Luc Richier, and Roland Groz. KameleonFuzz: Evolutionary fuzzing for black-box XSS detection. In *Proceedings of the 4th ACM Conference on Data and Application Security and Privacy - CODASPY '14*, pages 37–48, San Antonio, Texas, USA, 2014. ACM Press. http://dl.acm.org/citation.cfm?doid=2557547.2557550.

[14] F5, Inc. NGINX. https://www.nginx.com/, January 2021. Accessed: 2021-01-13.

[15] Michael Felderer, Matthias Büchler, Martin Johns, Achim D. Brucker, Ruth Breu, and Alexander Pretschner. Chapter one - security testing: A survey. volume 101 of *Advances in Computers*, pages 1 – 51. Elsevier, 2016.

[16] Seth Fogie, Jeremiah Grossman, Robert Hansen, Anton Rager, Petko Petkov, and an O'Reilly Media Company. Safari. XSS Attacks. https://go.oreilly.com/queensland-university-of-technology/library/view/-/9780080553405/?ar, 2011.

[17] Jose Fonseca, Marco Vieira, and Henrique Madeira. Testing and Comparing Web Vulnerability Scanning Tools for SQL Injection and XSS Attacks. In *13th Pacific Rim International Symposium on Dependable Computing (PRDC 2007)*, pages 365–372, Melbourne, Australia, December 2007. IEEE. http://ieeexplore.ieee.org/document/4459684/.

[18] Bernhard Garn, Ioannis Kapsalis, Dimitris E. Simos, and Severin Winkler. On the applicability of combinatorial testing to web application security testing: A case study. In *Proceedings of the 2014 Workshop on Joining AcadeMiA and Industry Contributions to Test Automation and Model-Based Testing - JAMAICA 2014*, pages 16–21, San Jose, CA, USA, 2014. ACM Press. http://dl.acm.org/citation.cfm?doid=2631890.2631894.

[19] Bernhard Garn, Marco Radavelli, Angelo Gargantini, Manuel Leithner, and Dimitris E. Simos. A Fault-Driven Combinatorial Process for Model Evolution in XSS Vulnerability Detection. In Franz Wotawa, Gerhard Friedrich, Ingo Pill, Roxane Koitz-Hristov, and Moonis Ali, editors, *Advances and Trends in Artificial Intelligence. From Theory to Practice*, volume 11606, pages 207–215. Springer International Publishing, Cham, 2019. http://link.springer.com/10.1007/978-3-030-22999-3_19.

[20] Shashank Gupta and B. B. Gupta. Cross-Site Scripting (XSS) attacks and defense mechanisms: Classification and state-of-the-art. *International Journal of System Assurance Engineering and Management*, 8(S1):512–530, January 2017. http://link.springer.com/10.1007/s13198-015-0376-0.

[21] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 281–296, Baltimore, MD, August 2018. USENIX Association.

[22] Hannes Holm and Mathias Ekstedt. Estimates on the effectiveness of web application firewalls against targeted attacks. *Information Management & Computer Security*, 21(4):250–265, January 2013. https://doi.org/10.1108/IMCS-11-2012-0064.

[23] Paweł Krawczyk and Heiderich Mario. HTML5sec-Injections-Jhaddix.txt. https://github.com/danielmiessler/SecLists/blob/bb915befb208fd900592bb5a25d0c5e4f869f8ea/Fuzzing/HTML5sec-Injections-Jhaddix.txt, 2019. Accessed: 2020-10-31.

[24] Qing Li, Jinfu Chen, Yongzhao Zhan, Chengying Mao, and Huanhuan Wang. Combinatorial Mutation Approach to Web Service Vulnerability Testing Based on SOAP Message Mutations. In *2012 IEEE Ninth International Conference on E-Business Engineering*, pages 156–162, Hangzhou, China, September 2012. IEEE. http://ieeexplore.ieee.org/document/6468233/.

[25] Mirakhorli, Mehdi and Galster, Matthias and Williams, Laurie. Understanding software security from design to deployment. *SIGSOFT Softw. Eng. Notes*, 45(2):25–26, April 2020.

[26] Mahmoud Mohammadi, Bill Chu, Heather Richter Lipford, and Emerson Murphy-Hill. Automatic web security unit testing: XSS vulnerability detection. In *Proceedings of the 11th International Workshop on Automation of Software Test - AST '16*, pages 78–84, Austin, Texas, 2016. ACM Press. http://dl.acm.org/citation.cfm?doid=2896921.2896929.

[27] Nabil M. Mohammed, Mahmood Niazi, Mohammad Alshayeb, and Sajjad Mahmood. Exploring software security approaches in software development lifecycle: A systematic mapping study. *Computer Standards & Interfaces*, 50:107 – 115, 2017.

[28] NBS System. Naxsi. https://github.com/nbs-system/naxsi, October 2020. Accessed: 2020-10-31.

[29] OWASP. FAQ – OWASP ModSecurity Core Rule Set. https://coreruleset.org/faq/, 2020. Accessed: 2020-11-08.

[30] OWASP. ModSecurity Core Rule Set. https://coreruleset.org/, 2020. Accessed: 2020-10-31.

[31] p0pr0ck5. P0pr0ck5/lua-resty-waf. https://github.com/p0pr0ck5/lua-resty-waf, October 2020. Accessed: 2020-10-31.

[32] PortSwigger Ltd. Burp Suite. https://portswigger.net/burp, January 2021. Accessed: 2021-01-13.

[33] B. Potter and G. McGraw. Software security testing. *IEEE Security Privacy*, 2(5):81–85, 2004.

[34] Rahman, Akond and Williams, Laurie. A bird's eye view of knowledge needs related to penetration testing. In *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security*, HotSoS '19, New York, NY, USA, 2019. Association for Computing Machinery.

[35] Kenneth Reitz. Requests: HTTP for Humans. https://requests.readthedocs.io/en/master/, 2020. Accessed: 2020-10-31.

[36] rsnake. XSS-RSNAKE.txt. https://github.com/danielmiessler/SecLists/blob/master/Fuzzing/XSS/XSS-RSNAKE.txt, 2019. Accessed: 2020-10-31.

[37] s7x. XSS-Cheat-Sheet-PortSwigger.txt. https://github.com/danielmiessler/SecLists/blob/master/Fuzzing/XSS/XSS-RSNAKE.txt, 2019. Accessed: 2020-10-31.

[38] Mike Samuel, Prateek Saxena, and Dawn Song. Context-sensitive auto-sanitization in web templating languages using type qualifiers. In *Proceedings of the 18th ACM Conference on Computer and Communications Security - CCS '11*, page 587, Chicago, Illinois, USA, 2011. ACM Press. http://dl.acm.org/citation.cfm?doid=2046707.2046775.

[39] Hossain Shahriar and Mohammad Zulkernine. MUTEC: Mutation-based testing of Cross Site Scripting. In *2009 ICSE Workshop on Software Engineering for Secure Systems*, pages 47–53, Vancouver, BC, Canada, May 2009. IEEE. http://ieeexplore.ieee.org/document/5068458/.

[40] D. E. Simos, R. Kuhn, A. G. Voyiatzis, and R. Kacker. Combinatorial Methods in Security Testing. *Computer*, 49(10):80–83, 2016.

[41] Dimitris E. Simos, Bernhard Garn, Jovan Zivanovic, and Manuel Leithner. Practical Combinatorial Testing for XSS Detection using Locally Optimized Attack Models. In *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 122–130, Xi'an, China, April 2019. IEEE. https://ieeexplore.ieee.org/document/8728914/.

[42] Jatesh Jagraj Singh, Hamman Samuel, and Pavol Zavarsky. Impact of Paranoia Levels on the Effectiveness of the ModSecurity Web Application Firewall. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 141–144, South Padre Island, TX, April 2018. IEEE. https://ieeexplore.ieee.org/document/8367754/.

[43] The Apache Software Foundation. Apache HTTP Server Project. https://httpd.apache.org/, January 2021. Accessed: 2021-01-13.

[44] Michael Thelin. Web Security: Cross-site scripting attacks using UTF-7. http://michaelthelin.se/security/2014/06/08/

web-security-cross-site-scripting-attacks-using-utf-7.html, 2014. Accessed: 2020-11-22.

[45] Omer Tripp, Omri Weisman, and Lotem Guy. Finding your way in the testing jungle: A learning approach to web security testing. In *Proceedings of the 2013 International Symposium on Software Testing and Analysis - ISSTA 2013*, page 347, Lugano, Switzerland, 2013. ACM Press. http://dl.acm.org/citation.cfm?doid=2483760.2483776.

[46] Trustwave Spiderlabs. ModSecurity: Open Source Web Application Firewall. https://modsecurity.org/, 2020. Accessed: 2020-10-31.

[47] I. A. Tøndel, M. G. Jaatun, and J. Jensen. Learning from software security testing. In *2008 IEEE International Conference on Software Testing Verification and Validation Workshop*, pages 286–294, 2008.

[48] Andrew van der Stock, Brian Glas, Neil Smithline, and Torsten Gigler. OWASP Top 10 2017. https://owasp.org/www-project-top-ten/2017/. Accessed: 2020-10-25.

[49] Michael Wagner, Kristoffer Kleine, Dimitris E. Simos, Rick Kuhn, and Raghu Kacker. CAGEN: A fast combinatorial test generation tool with support for constraints and higher-index arrays. In *2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 191–200, Porto, Portugal, October 2020. IEEE. https://ieeexplore.ieee.org/document/9155722/.

[50] Joel Weinberger, Prateek Saxena, Devdatta Akhawe, Matthew Finifter, Richard Shin, and Dawn Song. A Systematic Analysis of XSS Sanitization in Web Application Frameworks. In Vijay Atluri and Claudia Diaz, editors, *Computer Security – ESORICS 2011*, volume 6879, pages 150–171. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. http://link.springer.com/10.1007/978-3-642-23822-2_9.

[51] Yichun Zhang. OpenResty. https://openresty.org/en/, January 2021. Accessed: 2021-01-13.

[52] ZAP Dev Team. OWASP Zed Attack Proxy (ZAP). https://www.zaproxy.org/, January 2021. Accessed: 2021-01-13.

Modeling MCPTT and User Behavior in ns-3

Wesley Garey[1], Thomas R. Henderson[2], Yishen Sun[1], Richard Rouil[1], and Samantha Gamboa[3,4]

[1]Wireless Networks Division, National Institute of Standards and Technology, Gaithersburg, Maryland, United States
[2]Department of Electrical & Computer Engineering, University of Washington, Seattle, Washington, United States
[3]Associate, National Institute of Standards and Technology, Gaithersburg, Maryland, United States
[4]Prometheus Computing LLC, Sylva, North Carolina, United States

Abstract

To support the advancement of public safety communications technology, the Third Generation Partnership Project (3GPP) has created several standards to define Mission Critical Push-To-Talk (MCPTT) over Long Term Evolution (LTE) networks. As this is a new service that can be used in dire situations, it is imperative that the behavior and performance meet the expectations of first responders. This paper introduces an extension to the network simulator, ns-3, that models MCPTT and user Push-To-Talk (PTT) activity, so that researchers can gain insights and evaluate the performance of this service. In this paper we will describe MCPTT based on 3GPP definitions, the implementation of the MCPTT model in ns-3, and some results, including Key Performance Indicators (KPIs), that can be extracted from this model.

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 (NIST), nor is it intended to imply that the commercial products identified are necessarily the best available for the purpose.

1 INTRODUCTION

Land Mobile Radio (LMR) systems for two-way voice communications have been extensively used by public safety, first responder systems for decades. Many LMR systems (analog and digital) are presently deployed worldwide. However, in an effort to improve first responder communications, public safety agencies worldwide are adopting broadband cellular systems, i.e. 4G Long Term Evolution (LTE) and eventually 5G Systems (5GS). Besides offering new broadband data services and prioritized service to first responders, a key new service called Mission Critical Push-To-Talk (MCPTT) is being deployed, with the long term objective of replacing traditional LMR systems. MCPTT is based on a series of specifications by the Third Generation Partnership Project (3GPP) (3GPP, 2019), (3GPP, 2017a), (3GPP, 2017b).

From the user's perspective, LMR and MCPTT should offer a similar service experience, but the underlying technology supporting both (dedicated analog or digital radio channels, versus a shared, general-purpose, fixed-infrastructure cellular network) is significantly different. Successful voice communication despite possibly challenging network conditions is paramount for safety and mission effectiveness among first responders (Sun et al., 2019). Because of this, it is imperative that thorough studies are conducted to evaluate

Garey, Wesley; Henderson, Thomas; Sun, Yishen; Rouil, Richard A.; Gamboa Quintiliani, Samantha. "Modeling MCPTT and User Behavior in ns-3." Presented at 11th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (Simultech 2021), Paris, FR. July 07, 2021 - July 09, 2021.

the operation and performance of MCPTT. Field trials, lab testing, and interoperability exercises will be essential to fully vet MCPTT and instill confidence that it can be used to complement and possibly even replace LMR systems. High fidelity network simulations offer another framework to evaluate future MCPTT performance in ways that small-scale lab and field testing cannot easily accomplish. In particular, network simulators offer the ability to scale to large network sizes and public safety incident scenarios involving hundreds or even thousands of notional MCPTT users, and they allow for completely repeatable experiments.

This paper reports on what we believe to be the first publicly available simulation model for MCPTT, based on the network simulator, ns-3[1], whose 4G LTE models have been extended by several of our previous efforts (Rouil et al., 2017), (Gamboa et al., 2019), (Garey et al., 2020) to support models for 3GPP Proximity Services (ProSe), Device-to-Device (D2D) communications, and the User Equipment (UE) relay service, UE-to-Network Relay. Our MCPTT model provides standards-aligned implementations of the call control and floor control protocols defined by 3GPP, for both on-network and off-network operations, as well as traces to capture standardized Key Performance Indicator (KPI) measurements such as mouth-to-ear latency and network access time.

Section 2 gives some general information about MCPTT and its core components. Section 3 presents works related to this paper. Section 4 describes the ns-3 model, including its features, outputs, verification and limitations. Section 5 presents the reader with a case study to show how one can simulate MCPTT using our model. Finally, Section 6 provides concluding remarks.

2 BACKGROUND

A Push-To-Talk (PTT) service is used to provide an arbitrated method for two or more users to communicate. In terms of a simple voice communication service, at least from the user's perspective, this service could provide a means of communication that is very similar to that of what can be achieved with walkie-talkies. In the ideal case, when a user wants to talk, the user will request permission to do so, traditionally, by pushing the PTT button on their communication device (3GPP, 2019), and then the service is expected to handle this request by allocating resources and granting permissions accordingly to enable communication between the users. MCPTT is an advanced PTT service based on LTE that can be used by first responders during mission critical situations (3GPP, 2019). Thus, it is expected that MCPTT services will be available regardless of the user's connectivity with an LTE network (3GPP, 2019). This is why MCPTT supports two modes of operation: on-network and off-network.

On-network mode follows the traditional client-server architecture, where the PTT service is provided by a centralized MCPTT server via the internet that is accessible through the LTE core network. The off-network case follows the peer-to-peer architecture, with the PTT service being provided by the client devices in a distributed manner. This is made possible with the use of ProSe, which enables D2D communication for devices in close proximity. In this case the MCPTT client is responsible for allocating resources and performing arbitration. In either mode, it is expected that this service will be available, allow users to request permission to talk, and provide a deterministic mechanism to arbitrate between those requests (3GPP, 2019).

There are two main components that comprise MCPTT, call control and floor control. These components are defined respectively in 3GPP (2017a) and 3GPP (2017b). Several of the primary elements associated with each are also shown in Figure 1. Call control is responsible for managing call state coordination, including aspects such as resource allocation, management of logical channels, user information, group affiliation, and call type. Call control supports various types of calls, including basic group calls, private calls, and broadcast group calls. Group calls may be established for communication between two or more users in a particular MCPTT group, while private calls can exist between two users regardless of their group affiliation. Call control also supports basic, imminent peril, and emergency call types. These call types indicate the status of a call and are used to determine which network resources will be used by that particular call. Calls can be established, upgraded, downgraded, joined, left, rejoined, and released, and these actions are all accomplished through call control.

Floor control is responsible for providing the arbitration logic to determine who can talk at any given time during an ongoing call (3GPP, 2017b). MCPTT floor control currently supports queuing of requests, priority, and preemption. On-network floor control also supports dual speakers (i.e., two users talking at the same time). Queuing is a feature that can be used to coordinate passing the floor around from one requesting user to another during a busy call. Additionally, priority and preemption make it possible for users with a higher

[1] https://www.nsnam.org/

Garey, Wesley; Henderson, Thomas; Sun, Yishen; Rouil, Richard A.; Gamboa Quintiliani, Samantha. "Modeling MCPTT and User Behavior in ns-3." Presented at 11th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (Simultech 2021), Paris, FR. July 07, 2021 - July 09, 2021.

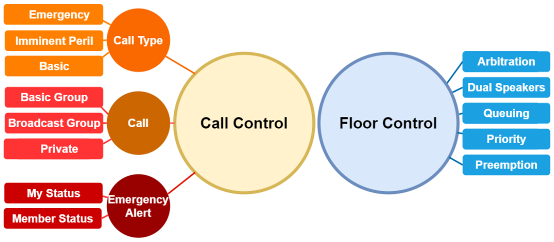


Figure 1: Main MCPTT components.

priority to interrupt the current speaker and take the floor immediately. While the assignment of user priority is not defined in the standard, it would be possible for users with a higher level of operational authority to always have higher priority than users with less authority, and to temporarily assign higher priorities to any user depending on their status (e.g., a user that ends up in an emergency situation).

Figure 2 illustrates some of the signaling that could take place during a basic, off-network, group call between two MCPTT users. The call control signaling is highlighted in orange while the floor control portion is highlighted in blue. In this example the call is initiated by a PTT push on Client 1 that triggers a probe to be sent four times, and if there is no response, the client concludes that there is no existing call for the group. After determining a call does not exist, Client 1 allocates floor control resources and sends out a call announcement to the other members in the group. Client 2 receives the announcement, accepts the call, and also allocates floor control resources. At this point, Client 1 grants itself the floor and notifies the user that they may begin speaking, at which point the client begins to send media packets that capture the user's voice to Client 2. During Client 1's transmission the user of Client 2 pushes the PTT button, which results in sending a floor request to Client 1. Since queuing is enabled in this example, Client 1 places the received request in the queue and responds with a queue information message to Client 2. Upon receiving the queue information message, Client 2 makes the user aware that the user must wait to speak. After the user of Client 1 is done speaking, the user releases the PTT button which results in Client 1 sending a granted message to Client 2, indicating that the user of Client 2 may begin speaking. Client 2 then makes the user aware, who then pushes the PTT button to accept the grant and begin speaking. Once the user of Client 2 is finished speaking, the user releases the PTT button and a floor release message is sent to Client 1 to indicate that the user of Client 2 is done speaking. At this point, the floor would be idle but could continue to be passed around between Client 1 and Client 2 until the call ends or is released by the users on the call. In addition, any of the features that were previously discussed, such as upgrading the call or additional users joining the call could also take place.

3 RELATED WORKS

In this section we will discuss some literature related to this model. First, this work extends our previously published work, including Sun et al. (2019) and Garey et al. (2020) that use earlier versions of the MCPTT model presented in this paper. Specifically, Sun et al. (2019) uses the ns-3 extension to create an analytical model that characterizes KPI 1 (access time) for off-network mode taking into account various factors that impact access time, including the scenario under which access is being requested, key ProSe and MCPTT

Garey, Wesley; Henderson, Thomas; Sun, Yishen; Rouil, Richard A.; Gamboa Quintiliani, Samantha. "Modeling MCPTT and User Behavior in ns-3." Presented at 11th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (Simultech 2021), Paris, FR. July 07, 2021 - July 09, 2021.

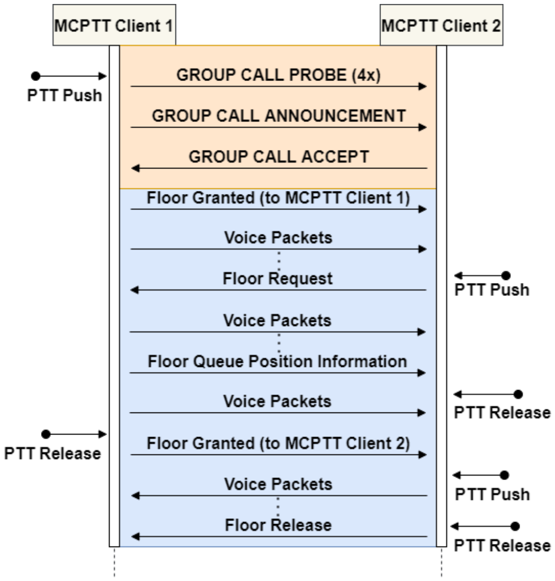


Figure 2: Off-network basic group call setup and use example.

configurations, and channel conditions. The work in Garey et al. (2020) is an extension of Sun et al. (2019) that also makes use of this model. In Garey et al. (2020), KPI 1, packet loss, delay, and data rate are used as performance indicators of the MCPTT application to study the impact of traffic, user density, range, and the underlying technology by considering both ProSe and Wi-Fi in the analysis. Our present work is distinguished by the inclusion of on-network MCPTT models, a novel orchestrated pusher model leveraging the translation of public safety call logs into empirical distributions driving the simulation, and a new example highlighting various modes of operation in a notional public safety scenario. Where our previous work in Gamboa et al. (2019) analyzed UE-to-Network Relay performance with an abstracted MCPTT model, we are able herein to run a highly complete simulation implementation of MCPTT.

Outside of our own work, several others have published work related to MCPTT operations and performance. In Feng and Li (2019), a conflict-adaptive back-off solution is proposed and analyzed using a Markov chain to investigate the performance of an enhanced version of the floor control protocol that could increase the likelihood of a successful floor request when multiple requests are initiated simultaneously. Brady and Roy (2020) perform an analysis with data collected from LMR and LTE systems in addition to numerical computations to compare how a transition from LMR to LTE would reduce wait times for first responders that wish to speak. Choi et al. (2019) uses a testbed to verify that with LTE MCPTT KPI measurements meet their respective requirements. Kim et al. (2019) uses an MCPTT test bed to investigate how a state-based uplink-scheduler can be used to reduce the latency of MCPTT signaling, in order to improve the overall performance of an MCPTT service. Kim et al. (2018) shows how the use of dynamic schedulers can be used to prioritize physical resources in real time for MCPTT users to increase download and upload speeds when the LTE network is congested. Sanchoyerto et al. (2019) predicts that MCPTT services will perform much better with the use of 5G instead of 4G because of the estimated reduction in latency across the network. Höyhtyä et al. (2018) discusses how prioritization of physical resources and rapidly deployable networks can be utilized by mission critical services in 5G applications. Solozabal et al. (2018) proposes a Mobile Edge Computing (MEC) architecture specifically for the use of MCPTT over 5G that could reduce call setup times and observed KPI values. Atanasov et al. (2020) also discusses the use of MEC to support mission critical services but without the use of Internet Protocol Multimedia Subsystem (IMS) that can provide improvements in performance.

All of the related works mentioned above address key issues facing MCPTT and other mission critical service operations and performance. However, many of the analyses focus solely on a single aspect of either on-network or off-network mode operation. The model that we will present gives researchers an opportunity to perform more comprehensive studies of MCPTT using the ns-3 simulator to analyze KPIs, behavior, and other performance metrics with a high degree of control and repeatability. This model could also be used to verify, study, and combine the related literature mentioned in this paper, as well as to perform large scale simulations that would otherwise require a large degree of coordination and many resources to execute using a real system or testbed equipment, if it is even possible. Furthermore, since our model is published as an open source extension of ns-3, we present the user with an opportunity to expand on current and future contributions made by the ns-3 and public safety communication research communities.

4 MODEL

In this section we discuss the MCPTT model that is implemented in ns-3. To start, we describe the application model in Section 4.1, which is an abstraction of the MCPTT service defined by 3GPP in 3GPP (2017a) and 3GPP (2017b). Next, we describe the pusher model that simulates user PTT activity using call log data in Section 4.2. Then, in Section 4.3 we discuss what outputs can be collected from our model, and how we verified our model's behavior in Section 4.4. Finally, in Section 4.5 we discuss the model's limitations.

4.1 Application Model

As mentioned in Section 2, MCPTT is a mission critical voice communication service originally designed to operate over LTE, specifically, for public safety. Moreover, MCPTT is a relatively small part of a much larger system. This means that for the most realistic and comprehensive model of MCPTT, one must also model LTE and support ProSe. This made ns-3 a prime candidate for our MCPTT model, as it is an open source network simulator that contains models for a number of communication technologies including LTE and ProSe (Rouil et al., 2017). It also supports the development of application layer models, primarily for generating traffic, which is where MCPTT fits into the network stack.

We also looked into using several other tools such as Vienna[2] and SimuLTE[3]. However, we found that Vienna is primarily a system and link layer simulator based on Monte Carlo simulations that does not model applications at the packet level. While SimuLTE is capable of modeling applications at the packet level through the use of "modules", we determined

[2] https://www.nt.tuwien.ac.at/research/mobile-communications/vccs/
[3] https://simulte.com

from experimentation that ns-3 offered greater control of application behavior and traffic for our purposes. SimuLTE also focuses on network-controlled D2D which requires UEs to be in-coverage (Nardini et al., 2018), whereas the D2D model in ns-3 supports in-coverage, out-of-coverage, and partial coverage situations.

To extend ns-3 with an MCPTT model, we created two new ns-3 applications that follow the MCPTT service defined in 3GPP (2017a) and 3GPP (2017b) very faithfully. They are the `McpttPttApp`, which represents the client application, and the `McpttServerApp`, which represents the server application. These two ns-3 applications coordinate several components that are necessary to provide many of the features described in Section 2. The client-side application, whose architecture is depicted in Figure 3, is the entity that offers a user Application Programming Interface (API) for MCPTT. This includes functions for pushing/releasing the PTT button, initiating/releasing calls, upgrading/downgrading calls, and switching between multiple calls. It also contains some information, such as the MCPTT user ID that would normally be stored in the Mission Critical Services (MCS) Management Object (MO) from 3GPP (2018a). The server-side application is intended to be installed on a server node and primarily communicates with the client application using a combination of MCPTT and Session Initiation Protocol (SIP) defined messages.

Both applications maintain a set of `McpttCall` objects that represent MCPTT calls. Only one call can be "selected" at a time by the client application, and this is the call that is affected by functions from the application's API. Each call contains an `McpttCallMachine` object, an `McpttFloorMachine` object, and channels for floor control and media messages. The channels used for call control messages reside in and are maintained by the MCPTT applications. The `McpttCallMachine` and `McpttFloorMachine` classes serve as interfaces that are specialized by concrete classes to allow the user to configure a call taking into account the variety of call types and modes of operation. For example, the two state machines needed to perform an off-network, basic group call would be the state machines from Section 10.2.1 in 3GPP (2017a) and Section 7.2.3 in 3GPP (2017b), which are implemented by the `McpttCallMachineGrpBasic` and `McpttOffNetworkFloorParticipant` classes in our model. Other parameters and configurations, such as group affiliation and resource allocation, is handled by the `McpttCallMachine`, while priority and preemption is handled by the `McpttFloorMachine`.

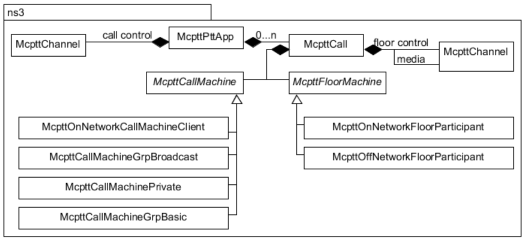


Figure 3: Client-side application architecture.

In addition to the main MCPTT components described above, we also created several supplementary components to facilitate MCPTT operations. This includes the `McpttCounter`, `McpttTimer`, `McpttChannel`, and `McpttFloorQueue` classes. `McpttCounter` and `McpttTimer` objects are used to model counters and timers, while the `McpttFloorQueue` class is used as needed to store floor requests when queuing is enabled. The `McpttChannel` class models the logical channels necessary for communication between the applications.

Along with those components we also model the call control and floor control messages necessary for signaling between various state machines, which also include a simplified

SIP module to handle the SIP related procedures and messages when operating in on-network mode. For example, the class `McpttFloorMsgRequest` represents the "Floor Request" message from the floor control protocol, and the `McpttMediaMsg` represents the Real-time Transport Protocol (RTP) media messages used to transmit voice when a user is speaking.

The application also includes a configurable media source to model the transmission of RTP packets when a user has the floor. This is accomplished by the `McpttMediaSrc` that can be configured to generate a payload of size S at a data rate of D. With this component the user of this model can mimic the traffic generated by a Constant Bit Rate (CBR) encoder. For example, the media source could be configured with $S = 60$ B and $D = 24$ kbit/s to send packets at a size and interval that resembles the traffic generated by the Adaptive Multi-Rate Wideband (AMR-WB) codec at 23.85 kbit/s (ITU-T, 2003).

4.2 Pusher Model

In addition to the protocol elements described in Section 4.1, we also model user activity to automate the use of the MCPTT service. This is accomplished by the `McpttPusher` class that is attached to each instance of an `McpttPttApp` object. We refer to this class as a "pusher" since it generates "push" and "release" events over the course of a simulation to mimic a user pushing and then later releasing the PTT button on their device. The default and simplest form of our pusher model is when an instance of the `McpttPusher` class acts independently, using random variables. In this mode of operation, which we refer to as "automatic mode," the pusher uses two random variables: one for generating a time span before a push indication will occur, and another for generating a time span before a subsequent release indication will occur.

Even though automatic mode allows the user to configure PTT activity for an individual pusher, it can be cumbersome to quantify, configure, and characterize the overall channel activity for a group of MCPTT users. This is due to the fact that PTT events are generated independently of one another in automatic mode. This also means that unlike the users of most public safety organizations, independent pushers do not take into account politeness or situational behavior based on what the other pushers' actions are or the current state of the call. As a result, we defined another approach, called "orchestrated mode," that provides a more coordinated pusher model across the users in a call. The benefits of orchestrated mode include the automated selection of pushers to produce PTT events and the ability to configure parameters at the system level according to the desired rate of activity. In contrast, automatic mode would require the ns-3 user to configure pushers by scaling the values used for each random variable based on the number of pushers that will be participating in each call to achieve an overall activity rate.

Orchestrated mode is achieved by using an instance of the `McpttPusherOrchestrator` class. This class is a centralized entity that controls a set of pushers by scheduling its own push and release events just like an individual pusher, but then randomly selects a pusher from the set it is orchestrating to carry out the actions. We also analyzed public safety call logs that were taken from an existing LMR system to further enhance the pusher model. During our analysis we determined that our model needed to capture two key elements to simulate the PTT behavior of such users: "talk spurts" and "talk sessions." We define the length of a talk spurt to be the time from when a PTT event begins until it ends, while the length of a talk session is simply the total length of consecutive talk spurts from one or more UEs in the same group. This resulted in the data sets plotted in Figure 4, for which a Cumulative Distribution Function (CDF) is used to capture the distribution of 8381 talk spurt lengths and 4606 talk session lengths. These data sets enable us to use orchestrated mode in combination with empirical random variables to resemble the overall PTT activity of a real public safety group call. The first enhancement comes from the `McpttPusherOrchestratorSpurtCdf` class, which creates and configures an underlying `McpttPusherOrchestrator` instance specifically to schedule talk spurt lengths based the talk spurt length CDF from Figure 4. The second enhancement comes from the `McpttPusherOrchestratorSessionCdf` class that can be attached to an `McpttPusherOrchestratorSpurtCdf` instance to enforce talk sessions based on the talk session length CDF in Figure 4.

While we were able to gather similar information about the interarrival times of talk spurts and talk sessions, we do not know if these statistics are highly dependent on the number of devices in a group call because while the call logs provide valuable information about active users, we do not know the system configuration, the total number of groups, or the size of those groups. Thus, we could not generalize interarrival times with respect to different group sizes. Therefore, we leave it up to the user to control interarrival times using an "activity factor." With this approach, we assume that the interarrival times of talk sessions and talk spurts follow an exponential distribution, which is scaled by the activity factor set by the user. This means that the user can specify individual values between (0, 1] for both parameters to determine the overall PTT activity during a simulation, with 0 representing no activity and 1 representing non-stop activity. The equation used to determine the mean interarrival time of talk spurts ($\overline{x_{IAT}}$) is:

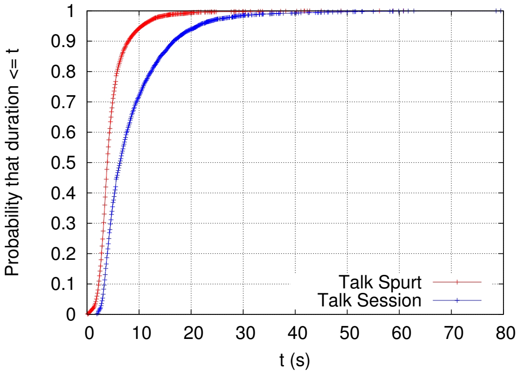


Figure 4: Talk spurt and talk session duration CDFs.

$$\overline{x_{IAT}} = \overline{x_T} * (\frac{1}{VAF} - 1) \tag{1}$$

where $\overline{x_T}$ is the average duration of a talk spurt, which we determined to be 4.69 s in our logs, and VAF is the activity factor set by the user. The equation used to determine the mean interarrival time of talk sessions ($\overline{y_{IAT}}$) is:

$$\overline{y_{IAT}} = \overline{y_T} * (\frac{1}{SAF} - 1) \tag{2}$$

where $\overline{y_T}$ is the average duration of a talk session, which we determined to be 8.58 s in our logs, and SAF is the activity factor set by the user.

For example, Figure 5 shows the resulting timeline of events if $VAF = 1$, $SAF = 0.25$, we have two pushers, and the total simulation time is 80 s. In this graph, the simulation time is on the x-axis while the states of sessions and pushers are on the y-axis. This indicates that there is an active talk session for 25 % of the total simulation time, and that within each active talk session a PTT button is pushed 100 % of the time. Note that, even though both VAF and SAF are configurable, setting $VAF = 1$ best matches the activity from the call logs that we analyzed. This is due to the fact that PTT events recorded in these logs and, consequently, captured in all of the statistics that we collected, include “hang time”. Hang time is an LMR system parameter that, as captured in the call logs, is additional time added to the end of a user’s transmission so that if a consecutive transmission occurs for that same user before a given amount of time has passed, it is recorded as one event. This means that it is likely the case that there were more occurrences of PTT events than what

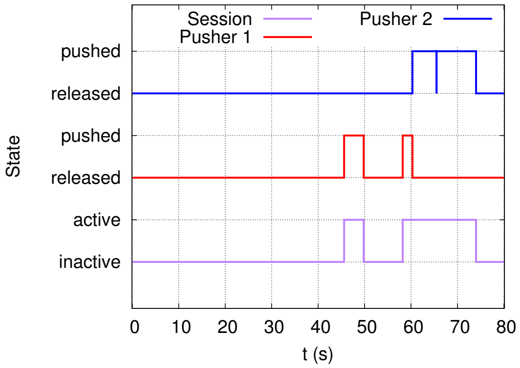


Figure 5: Orchestrated pusher model example.

is recorded in the logs, and the actual lengths of those PTT events were shorter. For example, if the system hang time is 3 s, a user pushes the PTT button for 2 s, releases the PTT button, and then 1 s later pushes the PTT button again for another 2 s before releasing it, this is recorded as one PTT event with a duration of $2+1+2+3=8$ s.

Our third enhancement to orchestrator mode comes from the `McpttPusherOrchestratorContention` class that accounts for overlapping PTT events, which are necessary for queuing and preemption to occur. An instance of this class can be attached to the `McpttPusherOrchestrator` to simulate colliding PTT events. This is accomplished by generating a value between [0, 1) with a uniform random variable and comparing it to a threshold set by the user whenever a PTT event scheduled by the underlying orchestrator occurs. We refer to this threshold as the Contention Probability (CP), and if the random value generated is less than CP then an additional PTT event will occur at anytime during the original event. For example, if there were originally 100 PTT events in the previous example, we then included an `McpttPusherOrchestratorContention` instance, and set $CP = 0.1$ (10 %), then approximately 10 out of the 100 PTT events generated by the original orchestrator would trigger an additional overlapping PTT event for a total of 110 PTT events. This means that, in total, approximately 9 % of all PTT events would be generated by the newly included orchestrator. In addition to the overall probability, to calculate the probability that any individual pusher's PTT event will overlap with another pusher's PTT event (PC_I), Equation 3 can be used.

$$PC_I = \frac{2*CP}{1+CP} \tag{3}$$

4.3 Outputs

The current implementation of our model contains several traces that can be used to derive information about various aspects of the MCPTT application. The message and state machine traces can be parsed post-simulation to study behavior and performance.

time (s)	nodeid	rx/tx	bytes	message
30.54	1	TX	58	McpttFloorMsgGranted
30.59	2	RX	58	McpttFloorMsgGranted

Table 1: Message trace snippet.

The message trace captures all MCPTT application message exchanges between state machines during a simulation. This includes all call control, floor control, and media messages, each of which can be filtered at the user's discretion. Table 1 shows the column names and two sample rows of data that can be included in this trace. Both traces include general information such as the ID of the node that created the entry in the trace and the time at which that record was created. Columns that are included specific to the message trace are those that indicate whether or not a message was sent or received, the size of the message, and the name of the message. For example, the first row of data in Table 1 indicates that 30.54 s into the simulation, node 1 sent a "Floor Granted" message that was 58 B in size. With this trace, information such as packet loss, delay, jitter, and data rate can be determined with varying granularity (e.g., per call, per application, etc.).

The data included in the state machine trace captures internal state information for various MCPTT state machines. Therefore, on top of the general fields, it also includes fields such as the name of the state machine that the record is for, the previous state of the state machine, and the new state of the state machine. With this trace, information such as the status of a call and the state of an application can be extracted. The user could also use this trace to detect operations and behavior that does not result in or is not the result of any message exchanges, such as inactivity or termination events.

In addition to the basic traces mentioned above, we also have traces to capture KPIs that measure the performance of MCPTT as defined by 3GPP (2019). The KPIs that our model is currently capable of reporting include access time (KPI 1) and mouth-to-ear latency (KPI 3).

To capture KPI 1, which measures the delay between a request to speak and the corresponding grant, our model uses the state transitions of the floor control state machine. The initial request to speak is indicated by a change to the 'O: Pending Request' state, and the grant is indicated by a subsequent transition to the 'O: Has Permission' state. The trace includes a result field to capture the outcome of a PTT request as there are several that need to be considered. The delay and outcomes of KPI 1 can be traced to an output file as depicted in Table 2.

time (s)	userid	callid	result	latency (s)
7.246	1	0	I	0.024
25.907	3	1	Q	2.653
28.952	4	1	I	0.024
34.017	3	1	D	0.651

Table 2: KPI 1 trace snippet.

The five possible outcomes that may result from a PTT request can be filtered based on the fourth column in Table 2, with "I" denoting "immediate", "Q" denoting "queued", "D" denoting "denied", "F" denoting "failed", and "A" denoting "abandoned". The meaning of each outcome value can be found in Table 3. Table 2 illustrates that 7.246 s into the simulation, the user with MCPTT User ID 1, was granted the floor immediately upon request and it took 24 ms. When measuring access time using this trace only "I" and "Q" outcomes align with the definition of KPI 1, and it is important to note that "Q" outcomes will be heavily dependent on pusher behavior. It is also worth mentioning that 3GPP (2019) specifies that KPI 1 should be less than 300 ms for 99 % of all on-network MCPTT requests when there is negligible backhaul delay and less than 70 % load per node. There is no similar requirement defined for off-network mode as of this writing.

Result	Definition
I	Granted immediately.
Q	Queued and subsequently granted.
D	Denied immediately.
F	Timeout or unexpected sequence.
A	PTT button was released.

Table 3: KPI 1 trace outcome definitions.

To capture KPI 3 in our model, the RTP packets of a talk spurt include a timestamp to mark when they were generated. When those RTP packets are received by a receiving MCPTT application it will check whether the packet contains a newer 'start-of-talkspurt' timestamp, which indicates that a new talk spurt exists. The latency of each new talk spurt is then traced by each receiving application as depicted in Table 4. From the row in Table 4 we can see that 3.727 s into the simulation, node 8 detected a new talk spurt from Synchronization Source (SSRC) 3 that had a latency (i.e., KPI 3) of 24 ms.

time (s)	ssrc	nodeid	callid	latency (s)
3.727	3	8	1	0.024

Table 4: KPI 3 trace snippet.

4.4 Verification

MCPTT off-network models were subject to an extensive verification, previously published in Varin et al. (2018), against 3GPP (2017a) and 3GPP (2017b) to verify the model's behavior. We have verified and validated our on-network models in several ways. First, 3GPP defined conformance tests for selected on-network mode configurations in 3GPP (2020), and we defined ns-3 unit tests to align with the testing steps in that document. Second, Nemergent Solutions, a vendor based in Spain that develops Mission Critical solutions, has published detailed packet traces for MCPTT version 13.3 on-demand, on-network, pre-arranged group call with automatic commencement[4], and we aligned our basic tests with this trace. Finally, the Public Safety Communication Research (PSCR) division at the National Institute of Standards & Technology (NIST) conducted small-scale measurements of MCPTT on-network operations using a testbed and shared packet traces with us, with which we confirmed close alignment.

[4] https://nemergent.com/traces.html

4.5 Limitations

While we intend for our model to capture the behavior of an actual MCPTT system, it is not an actual system with real users. This comes with limitations that have the potential to affect studies performed with our model and may lead to unrealistic observations. The limitations that we will discuss in the remainder of this section include any components, features, and functionalities that we know to be lacking or absent in our current ns-3 implementation.

We mentioned in Section 4.1 that MCPTT is a part of a much larger system, and we must therefore also consider the limitations of the models that surround our MCPTT model. This includes limitations in ns-3's LTE model since it is the network technology that MCPTT was initially designed to operate over. The ProSe model does not support ProSe Per Packet Priority (PPPP), affecting the priority of physical resources used by an off-network MCPTT application when the call is in an elevated status. A similar limitation exists for on-network operation, since the LTE schedulers in ns-3 are not designed to prioritize on a bearer basis or within a bearer based on call status as specified by 3GPP (2017a). This means that any studies pertaining to call priority and resource allocation based on call status would require further development of the underlying ns-3 LTE model. Also, the UE-to-Network Relay model lacks coordination between the uplink (UL) and sidelink (SL) resource scheduling, which impacts the performance of any application running over relay (Gamboa et al., 2019). The ns-3 Evolved Packet Core (EPC) model in LTE is also simplistic, with idealized representations of the MME and no modeling of delays that may come from authorization, access control, roaming, server processing, etc. These gaps in modeling are not fundamental limitations but arise from the current state of the models. MCPTT-based simulation studies would benefit from future improvements to the ns-3 cellular modeling fidelity because prioritization and non-ideal control channels have the potential to affect measured KPIs and general statistics such as packet delay.

In Section 4.2 we already mentioned several factors from the call logs that impact our pusher model, but we did not mention additional studies that could be performed to enhance this model. For example, our model does not take into account user impatience and the effect of network congestion that leads to abandoned transactions and more user impatience, such as is done in Baynat et al. (2015). Such user behavior could also be directly studied with regard to public safety users in cases where PTT requests do not go through, and possibly influenced by different operational contexts, such as a natural disaster or a routine traffic stop. With that said, our model could be expanded to take into account environmental and/or situational details to realize more realistic PTT activities that could ultimately lead to more realistic traffic patterns.

5 CASE STUDY

At the annual Public Safety Broadband Stakeholder meeting, we demonstrated the use of this model in a large-scale, public safety, scenario with a notional duration of four hours and involving over one hundred nodes in various operational roles[5]. However, we focus herein on a smaller scale scenario because it more clearly highlights the key modes of operation (on-network via basic LTE or UE-to-Network Relay, and off-network via ProSe) that are currently supported. This scenario will be included in the next release of our public safety extensions for ns-3, with the basic goal of comparing performance metrics for three public safety network configurations.

This scenario consists of three teams of MCPTT users, with each team participating in its own MCPTT group call. All team members of team one are located within coverage of an eNodeB so they are all connected to the network (on-network). The second team is only connected with each other via ProSe since they are inside a building whose walls prevent connection to an eNodeB on the outside that would allow this team to connect to the network (off-network). A third team operates in a hybrid mode; all team members are within the building, but one team member is situated at the door with connectivity to both the team inside and the eNodeB (relay). This team member runs a UE-to-Network Relay service (defined in 3GPP (2018b)), connecting oneself and team members on the inside with the network. Figure 6 illustrates the scenario topology. The simulation can be scaled to different team sizes (with a default of 4 users per team) and run with a configurable simulation duration.

The ns-3 simulator has abstractions for the physical layer models, allowing insertion of path loss models that affect the packet error rates that are experienced in a simulation. This includes models that take building wall penetrations into account, one of which, the `HybridBuildingsPropagationLossModel`, we use in this example. The network technology

[5] https://www.nist.gov/ctl/pscr/simulation-and-visualization-public-safety-incidents

Garey, Wesley; Henderson, Thomas; Sun, Yishen; Rouil, Richard A.; Gamboa Quintiliani, Samantha. "Modeling MCPTT and User Behavior in ns-3." Presented at 11th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (Simultech 2021), Paris, FR. July 07, 2021 - July 09, 2021.

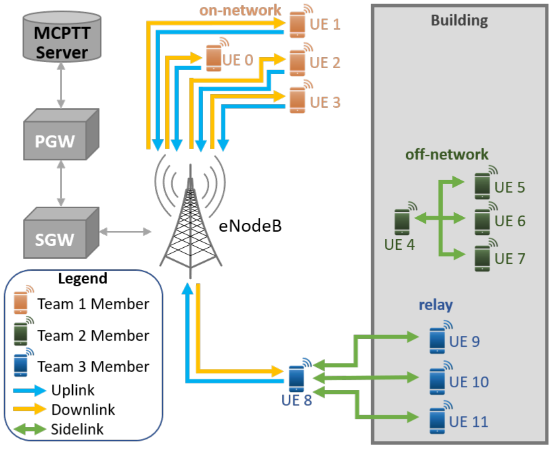
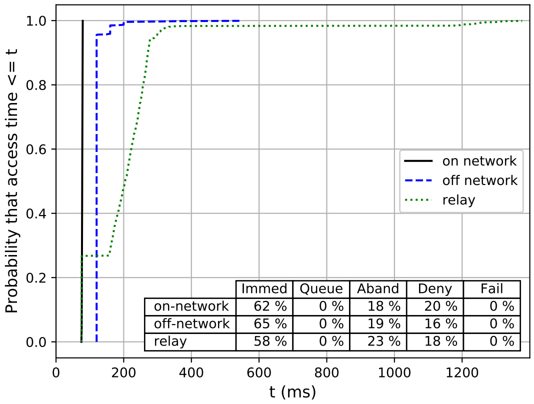


Figure 6: Example MCPTT scenario

is LTE, with a notional EPC network and a public safety server that houses the MCPTT server. In this example, the EPC round-trip delay for on-network and relay operation (from eNodeB to the MCPTT server, and back) is configured as a fixed 60 ms. The off-network SL period is configured to its lowest possible value of 40 ms.

We highlight performance here by post-processing the raw KPIs described in Section 4.3 to generate empirical CDFs on a per-team basis. This includes the access time performance (KPI 1) and the mouth-to-ear latency performance (KPI 3) that are observable once an MCPTT user has successfully obtained access to the floor. Each team is configured with automatic pushers governed by normal random variables with a mean IAT of 10 s, and a mean duration of 1 s (negative values are truncated to zero). Therefore, the overall pusher busy time is roughly $\frac{4}{11}$ or 36 %.

The first set of results presented is from a configuration that disables queuing of floor requests, so any request for the floor will be either granted immediately, denied, or fail or be abandoned if the floor request is not served in time. Figure 7 plots the empirical CDF of the access times observed for floor requests, for a simulation that runs until at least 1000 successful network accesses are observed for each team. Only values for immediately served or successfully queued requests contribute to the CDF of access times. Also tabulated, for each team, are the percentages observed for each type of outcome.

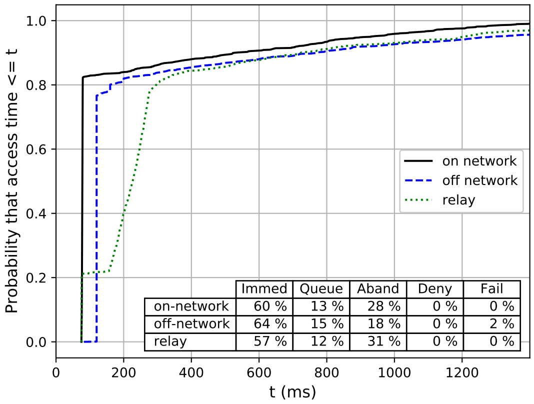


	Immed	Queue	Aband	Deny	Fail
on-network	62 %	0 %	18 %	20 %	0 %
off-network	65 %	0 %	19 %	16 %	0 %
relay	58 %	0 %	23 %	18 %	0 %

Figure 7: MCPTT Access Time for immediately granted requests with queuing disabled.

On-network access time is expected to provide the lowest minimum latency, because the server is centralized and can determine access immediately. In this first case without queuing enabled, on-network access time represents the sum of the core network delay (60 ms, as stated above), and the Radio Access Network (RAN) delay, which ranges from 15 ms to 19 ms in this configuration, leading to a nearly vertical line in the plot at around 75 ms. For the off-network team, access time should be around 120 ms most of the time, when the floor is idle, because by default the distributed protocol uses 3 SL periods to be confident that the floor is available before taking it. However, it could take longer because the floor control protocol also takes into account when multiple users are requesting the floor at the same time, at which point, additional SL periods may be required. In the illustrated case, it takes more time in off-network mode for floor control coordination than the round-trip time it takes for the request/response exchange with the on-network server. The UE-to-Network Relay team's performance in this case exceeds that of the on-network case, due to the extra relay hop introducing additional delay and loss in about 75 % of the requests. The lowest 25 % of values in the relay case correspond to floor requests initiated by the relay team member itself, which do not require any SL transmissions. As explained in more detail in Gamboa et al. (2019), message loss can happen due to scheduling decisions at the relay (prioritizing UL over SL transmissions, serving the attached nodes using round robin, and SL period transmission cutoff) and due to the half-duplex operation of SL transmission, so message loss may be significantly higher as compared to the on-network case. The CDF points that appear as extreme outliers of latency (greater than 1 s) are due to floor request retransmissions, the timer for which has a timeout value of 1 second.

Figure 8 shows the access time CDF from a similarly configured simulation, with the only difference being that queuing of floor requests is enabled for both on-network and



	Immed	Queue	Aband	Deny	Fail
on-network	60 %	13 %	28 %	0 %	0 %
off-network	64 %	15 %	18 %	0 %	2 %
relay	57 %	12 %	31 %	0 %	0 %

Figure 8: MCPTT Access Time for queued or immediately granted requests with queuing enabled.

off-network. In general, queuing increases the probability that a request will be served, at the expense of additional access time delay. If the floor request is queued, then the time to wait for the current talker, or those earlier in the queue, will be added to the latency. The plot shows that roughly 80 % of the values are immediately served, and the remaining 20 % of the values are queued (abandoned, denied, and failed requests are not counted in the CDF). The CDF slope above 80 % is gradual and reaches maximum values of between 2 s and 4 s for the three teams; these latencies are dependent on waiting for one or more talk spurts to end.

For this scenario, the plot shown in Figure 9 illustrates an empirical CDF of the mouth-to-ear latency of the beginning of each talk spurt in the same simulation scenario with queuing disabled. Unlike the access time statistics, the mouth-to-ear latency statistics are unaffected by the queuing configuration because the latency samples are taken after the floor has been granted. This CDF illustrates that with off-network operation, the RTP packets can flow directly between devices and do not have to go to the eNodeB, core network, server, and then back. On-network takes longer, as expected, because of the round trip that RTP packets incur (to the MCPTT server and back), while the relay experiences even more latency, as expected, due to the additional SL hop that introduces higher delays and loss. When queuing is disabled, the on-network mouth-to-ear latency and the access time latency are roughly the same (one RTT), as shown. One difference between Figures 7 and 9 for relay communications is that we no longer observe latency greater than 1 s for the data traffic because there are no lost floor request messages that need to be re-transmitted.

Garey, Wesley; Henderson, Thomas; Sun, Yishen; Rouil, Richard A.; Gamboa Quintiliani, Samantha. "Modeling MCPTT and User Behavior in ns-3." Presented at 11th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (Simultech 2021), Paris, FR. July 07, 2021 - July 09, 2021.

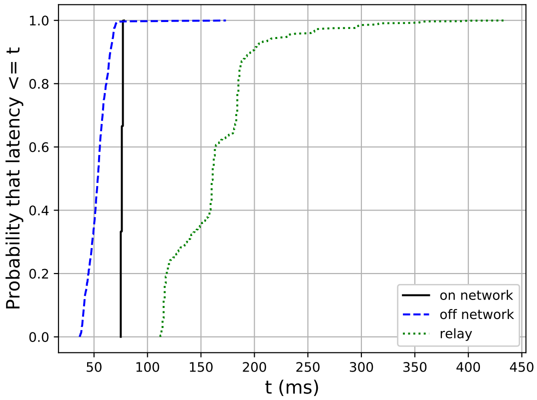


Figure 9: MCPTT Mouth-to-Ear Latency.

6 CONCLUSION

In this paper we presented an MCPTT simulation model that extends the network simulator, ns-3. Based on our earlier work to develop standards-aligned call control and floor control simulation models for MCPTT off-network operation, we describe herein the extension to on-network operation, the enhancement of our PTT traffic generation model (based on distilling PTT call traces to parameterize stochastic talk events in ns-3), and a public safety example scenario. We also summarize the basic operation of MCPTT in practice, the architecture of the ns-3 MCPTT model and limitations thereof, the output data available for analysis, and how we verified the model using test cases and published traces from an on-network MCPTT testbed. Future work will include the development of new and existing components, features, and functionalities discussed in Section 4.5 that will enhance the accuracy and behavior of our current MCPTT model and the models that surround MCPTT for a more realistic representation of the service. Future work will also include the use of this model to further analyze and develop public safety scenarios to study the impact of system configurations, traffic loads, etc. on MCPTT performance.

REFERENCES

3GPP (2017a). Mission Critical Push To Talk (MCPTT) call control; Protocol specification. Technical Specification 24.379, 3rd Generation Partnership Project (3GPP). Version 14.4.0.

3GPP (2017b). Mission Critical Push To Talk (MCPTT) media plane control; Protocol specification. Technical Specification 24.380, 3rd Generation Partnership Project (3GPP). Version 14.4.0.

3GPP (2018a). Mission Critical Services (MCS) Management Object (MO). Technical Specification 24.483, 3rd Generation Partnership Project (3GPP). Version 14.4.0.

3GPP (2018b). Proximity-based services (ProSe); Stage 2. Technical Specification 23.303, 3rd Generation Partnership Project (3GPP). Version 15.1.0.

3GPP (2019). Mission Critical Push To Talk (MCPTT); Stage 1. Technical Specification 22.179, 3rd Generation Partnership Project (3GPP). Version 16.5.0.

3GPP (2020). Mission Critical (MC) services over LTE; Part 2: Mission Critical Push To Talk (MCPTT) User Equipment (UE) Protocol conformance specification. Technical Specification 36.579-2, 3rd Generation Partnership Project (3GPP). Version 14.6.0.

Atanasov, I., Pencheva, E., and Nametkov, A. (2020). Handling Mission Critical Calls at the Network Edge. In *2020 International Conference on Mathematics and Computers in Science and Engineering (MACISE)*, pages 6–9.

Baynat, B., Vasseur, M., and Abreu, T. (2015). Revisiting the Characterization and the Modeling of User Impatience in Ubiquitous Networks. *PE-WASUN '15: Proceedings of the 12th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor*, pages 85–91.

Brady, C. and Roy, S. (2020). Analysis of Mission Critical Push-to-Talk (MCPTT) Services Over Public Safety Networks. *IEEE Wireless Communications Letters*, 9(9):1462–1466.

Choi, S. W., Song, Y., Shin, W., and Kim, J. (2019). A Feasibility Study on Mission-Critical Push-to-Talk: Standards and Implementation Perspectives. *IEEE Communications Magazine*, 57(2):81–87.

Feng, S. and Li, H. (2019). Floor Control Conflict Resolution in Off-network Mode of MCPTT. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 509–513.

Gamboa, S., Thanigaivel, R., and Rouil, R. (2019). System Level Evaluation of UE-to-Network Relays in D2D-Enabled LTE Networks. In *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–7.

Garey, W., Sun, Y., and Rouil, R. (2020). Performance Evaluation of Proximity Services and Wi-Fi for Public Safety Mission Critical Voice Application. *Wireless Communications and Mobile Computing*, 2020(8198767).

Höyhtyä, M., Lähetkangas, K., Suomalainen, J., Hoppari, M., Kujanpää, K., Trung Ngo, K., Kippola, T., Heikkilä, M., Posti, H., Mäki, J., Savunen, T., Hulkkonen, A., and Kokkinen, H. (2018). Critical Communications Over Mobile Operators' Networks: 5G Use Cases Enabled by Licensed Spectrum Sharing, Network Slicing and QoS Control. *IEEE Access*, 6:73572–73582.

ITU-T (2003). Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB). Technical Specification G.722.2, TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU (ITU-T). Version 16.5.0.

Kim, H., Jo, J., Park, C., Ahn, S., Chin, H., Park, P., and Kim, Y. (2018). Dynamic Resource Scheduling Algorithm for Public Safety Network. In *2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim)*, pages 127–132.

Kim, J., Jo, O., and Choi, S. W. (2019). State-Based Uplink-Scheduling Scheme for Reducing Control Plane Latency of MCPTT Services. *IEEE Systems Journal*, 13(3):2547–2550.

Nardini, G., Virdis, A., and Stea, G. (2018). Modeling Network-Controlled Device-to-Device Communications in SimuLTE. *Sensors (Basel, Switzerland)*, 18(10):3551.

Rouil, R., Cintrón, F., Ben Mosbah, A., and Gamboa, S. (2017). Implementation and Validation of an LTE D2D Model for Ns-3. In *Proceedings of the Workshop on Ns-3*, WNS3 '17, page 55–62, New York, NY, USA. Association for Computing Machinery.

Sanchoyerto, A., Solozabal, R., Blanco, B., and Liberal, F. (2019). Analysis of the Impact of the Evolution Toward 5G Architectures on Mission Critical Push-to-Talk Services. *IEEE Access*, 7:115052–115061.

Solozabal, R., Sanchoyerto, A., Atxutegi, E., Blanco, B., Fajardo, J. O., and Liberal, F. (2018). Exploitation of Mobile Edge Computing in 5G Distributed Mission-Critical Push-to-Talk Service Deployment. *IEEE Access*, 6:37665–37675.

Sun, Y., Garey, W., Rouil, R., and Varin, P. (2019). Access Time Analysis of MCPTT Off-Network Mode over LTE. *Wireless Communications and Mobile Computing*, 2019(2729370).

Varin, P., Sun, Y., and Garey, W. (2018). Test Scenarios for Mission Critical Push-To-Talk (MCPTT) Off-Network Mode Protocols Implementation. Technical Report 8236, NIST.

What Futuristic Technology Means for First Responders: Voices from the Field

Shaneé Dawkins[0000-0002-8114-0608], Kerrianne Morrison[0000-0002-2735-8809], Yee-Yin Choong[0000-0002-3889-6047], and Kristen Greene[0000-0001-7034-3672]

National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
shanee.dawkins@nist.gov, kerrianne.morrison@nist.gov,
yee-yin.choong@nist.gov, kristen.greene@nist.gov

Abstract. The public safety communication technology landscape in the United States (U.S.) is evolving to supplement the use of land mobile radios with a broader spectrum of communication technologies for use on the newly created Nationwide Public Safety Broadband Network. The goal of the multi-phase research study presented here was to understand the use of communication technologies by the population of first responders— Communications (Comm) Center & 9-1-1 Services; Emergency Medical Services; Fire Services; and Law Enforcement. The sequential, exploratory mixed methods study consisted of an initial exploratory qualitative phase followed by a larger quantitative phase. The qualitative data collection was via in-depth interviews with 193 first responders across the U.S.; the quantitative survey was completed by 7,182 first responders across the U.S. This paper presents the results of the study related to first responders' perceptions about the future of public safety communication technology. Discussed are the technologies first responders think would benefit their individual user populations, as well as communication technologies that would be useful across user populations within the public safety domain. Results show that first responders are open to new and exciting technologies, but their needs are utility driven; to have the biggest impact, their communication technology must be tailored to their needs and contexts. This paper will present the needs of first responders, in their own voices, to aid in the research and development of public safety communication technology.

Keywords: Usability, User Survey, UX (User Experience), User Requirements, Public Safety, First Responders, Incident Response.

1 Introduction

The public safety communication technology landscape in the United States (U.S.) is evolving. With the newly created Nationwide Public Safety Broadband Network (NPSBN), the public safety community is supplementing the use of land mobile radios with a broader spectrum of communication technologies. The public safety community has identified User Interfaces and User Experiences (UI/UX) as one of the key areas for research and development of these rapidly advancing technologies [1]. As such, the

Dawkins, Shanee; Buchanan, Kerrianne; Choong, Yee-Yin; Greene, Kristen K. "What Futuristic Technology Means for First Responders: Voices from the Field." Presented at 23rd International Conference on Human-Computer Interaction (HCI International 2021). July 24, 2021 - July 29, 2021.

Public Safety Communications Research (PSCR) Program at the National Institute of Standards and Technology (NIST) conducts research focusing on the end users – first responders [2]. Under this program, the NIST PSCR Usability Team performs research and provides guidance to ensure that communication technology in the public safety domain helps first responders achieve their goals and objectives with effectiveness, efficiency, and satisfaction in their specified contexts of use [3]. To this end, the NIST PSCR Usability Team has studied the public safety field to gain a better understanding of the user population of first responders — Comm Center & 9-1-1 Services (COMMS); Emergency Medical Services (EMS); Fire Services (FF); and Law Enforcement (LE). These four first responder disciplines, COMMS, EMS, FF, and LE, use different types of tools for different purposes; they experience different problems and have different communication technology needs. This is why it is crucial to understand the different public safety user groups and the communication technology they currently use, the problems they experience with current technology, and the technology they would like to have access to in the future.

NIST's PSCR Usability Team conducted a multi-phase, mixed methods research project in order to provide greater understanding of first responders, their experiences, and their communication technology problems and needs. The goal was to understand what first responders believe is necessary to facilitate communication and address their communication technology needs. Phase 1 of the project was a qualitative examination of first responder contexts of work [4]; interviews were conducted with first responders across the country from the four first responder disciplines—COMMS, EMS, FF, and LE. Phase 2 of the project utilized data from the qualitative interviews conducted in Phase 1 to create a large-scale, nationwide survey. The Phase 2 survey was designed to augment understanding of the types of communication technology first responders have, use, and want, and the problems they currently experience with their technology [5]. Understanding the use of communication technology by the four disciplines is critical to the success of the technology developed for the NPSBN.

Given the breadth and depth of the data collected, this paper focuses on a subset of the results from the Phase 2 survey related to the future of public safety and the NPSBN, presenting Phase 1 interview data throughout as appropriate. Previously analyzed results from the study are presented in [4 - 11]; additional data will be examined in future publications. The forward-looking communication technology needs of first responders presented here specifically focus on the potential usefulness of current devices first responders do not have, futuristic devices, and virtual reality (VR).

2 Methodology

2.1 Overview

The project consisted of a study with a sequential, exploratory mixed methods design, where an initial exploratory qualitative phase was followed by a larger quantitative phase. Phase 1 – the qualitative phase – examined first responders' communication technology use via in-depth interviews [4, 7]. The data from the interviews were the basis of the survey design used in Phase 2 – the quantitative phase [5]. In Phase 2, a large-scale, nationwide survey was conducted in order to gain a more comprehensive understanding of communication technology in the public safety community [5]. Data from both phases was integrated for analysis to provide for a more holistic understanding of first responders and their communication. For ease of exposition, this paper will refer to the research phases as “interviews” and “survey” henceforth (for the Phase 1 qualitative interviews and the Phase 2 quantitative survey, respectively).

Overarching Sampling Goals. To provide a representative sample of first responders in the U.S., multiple variables were considered to develop the sampling strategy in both phases of the study. The sampling strategy included first responders in a variety of positions within the four public safety disciplines – COMMS, EMS, FF, and LE. Due to the varied public safety issues faced in different parts of the country, geographic and cultural diversity were also primary considerations. Across the U.S., urban (U), suburban (S), and rural (R) districts were sampled to ensure that cities and districts of different sizes and different economic realities were represented. Another consideration was jurisdictional diversity, including federal, state, county and local jurisdictions; however, local jurisdictions had higher priority, as incident response typically starts at the local level. Other variables considered in the sampling strategy were career and volunteer FF, public and private EMS, and civilian and deputized COMMS. With the wide range of different types of first responders, their roles and responsibilities, and their different communication and technology needs, this approach provided insight into the many different experiences of public safety communication across the U.S., ensuring coverage of both typical and unique experiences.

The NIST Research Protections Office reviewed the protocol for this project and determined it met the criteria for “exempt human subjects research” as defined in 15 CFR 27, the Common Rule for the Protection of Human Subjects.

2.2 Interview methodology

The interviews were conducted with first responders across the U.S. in 2017 and 2018. There were three research questions:

1. How do public safety personnel describe the context of their work, including their roles and responsibilities as well as process and flow?
2. How do public safety personnel describe their communication and technology needs related to work?
3. What do public safety personnel believe is working or not working in their current operational environment related to communication and technology?

These research questions guided the interview protocol design and analysis, as [4] extensively reported.

Interview Sampling. Since demographic factors such as age, years of service, and gender may play a role in participants' views related to public safety communication, purposive sampling was applied in Phase 1. The sampling involved seeking participants who represented the full range of first responder experiences, as previously mentioned. Areas for in-person interviews were chosen that provided reasonable coverage of the depth and breadth of geographic and cultural diversity in the U.S., as well as the broad types of incidents that first responders face, aligning with eight of the ten U.S. Federal Emergency Management Agency (FEMA) regions [12].

Data Collection and Analysis. The data collection and analysis followed a rigorous qualitative research process. First, the in-depth interviews with first responders in the COMMS, EMS, FF, and LE disciplines were conducted in 45-minute sessions with first responders at their convenience (typically one-one-one at their station or department). These interviews were then audio recorded and transcribed. Two code lists were generated in order to label, or tag, participant statements: one for EMS, FF and LE, and one for COMMS, given the unique environment and primary tasks within that discipline. Then, the transcripts were coded according to the code lists, and the data were extracted (i.e., the data associated with a code from each transcript was exported into a separate document). Finally, themes were identified; relationships were examined among the codes, and between and within the four disciplines. This iterative process facilitated the identification of themes, trends, and outliers and provided an overall impression and understanding of the data. The themes, along with communication technology problems and needs findings, were used as the basis for the survey design in the second phase of the study.

2.3 Survey methodology

The survey development began at the conclusion of the interviews; Greene, et. al. extensively reported details about the survey instrument and survey methodology [5]. The following research questions served as guides for the development of the survey.

1. What are first responder needs related to communication and technology as they engage in their user-identified primary tasks?
   a. What communication tools and technology do first responders believe currently work, or do not work, for them?
2. What are the problems that first responders experience as they use communication technology?

The survey collected a wide variety of data related to communication technology use by first responders, from their day-to-day technology use, problems, and needs, to the technology that would be more suitable for use in larger, out of the ordinary incidents. Survey questions and response options were grounded in research from the previously collected empirical interview data, as well as from content and survey expert reviews during survey development. One of the driving ideals in the design of the survey was to keep it short out of respect for first responders and their time, and to encourage survey completion.

Survey Instrument Design. After a rigorous design process that included content and survey expert reviews, and given the myriad of different types of communication technology utilized and needed for the individual disciplines, it became clear there would need to be four different surveys, tailored for each discipline. The overall survey structure and flow were largely similar across the four survey versions: all began with a section on demographics, followed by a section on use of technology for day-to-day incident response (including questions on applications/software), and concluded with a section on use of technology in large events. The survey questions for EMS, FF, and LE were nearly identical, while differing somewhat more for COMMS, due to the different nature of their working environment [9]; for example, COMMS respondents were asked questions about call centers and Next Generation 9-1-1 (NG 911) [13, 14]. For all four disciplines, lists of technologies were used for questions about responders' use of day-to-day devices and devices used for large events. The lists of technologies used in the survey were catered to each discipline as the result of a thorough review of the problems and requested functionality identified in the interviews [7]. The goal was to not have first responders go through questions or lists of technologies that did not pertain to their work, as part of the effort to keep the survey short out of respect for first responders and their time. Greene et. al. reported detailed descriptions of survey logic, branching, and all questions and response options [5].

As this paper focuses on a subset of the survey data, the remainder of this section describes the details of the survey design solely related to the questions from which results are discussed. These questions, related to the potential usefulness of futuristic technologies for day-to-day incident response, are: 1) futuristic technologies; 2) NG 911 (COMMS only), and 3) VR.

Futuristic Technologies Question. The futuristic technology question was framed with the text, "We know there is no such thing as a "typical" day in public safety. However, for this set of questions, focus on the kinds of things you use in your day-to-day work." The question stem was "Which of the items below **would also be useful** for your **DAY-TO-DAY** work." Respondents were presented with a list of technologies and asked to "Check all that apply." The goal here was solely to identify those items that respondents believed would be useful in day-to-day incident response, not to have them rank these items or indicate whether they were more or less useful than other items.

The list of technologies in this question was populated from two sources. The first source was a preset list of technology based on PSCR research priorities and derived from the results of the interviews. Note that as previously mentioned, different first responder disciplines saw different lists of futuristic technologies, because the survey was driven by the interview data and the technologies that first responders discussed as potentially important for their work. The second source was a list of items that were piped forward based on a participant's previous survey responses about their day-to-day technology use. On a previous question, participants were asked about how often or not they use existing technologies. Every device for which they made no selection or selected "do not have" was piped forward to the future technology list. The items that were piped forward allowed respondents to select items they thought would be useful even if they did not currently have them.

In addition to the "Check all that apply" question, respondents were also provided with an open-ended text box where they could list additional technologies they thought would be useful or provide additional information.

Next Generation 9-1-1 Question. NG 911 is a digital or Internet Protocol (IP)-based 9-1-1 system that has several key capabilities, including: the ability for voice, photos, videos and text messages to be sent from the public to the 9-1-1 network; the transfer of emergency calls, location information, and multimedia to another PSAP; and the exchange of voice and data with other state or federal entities involved in the response via internetworking technologies based on open standards [13][14]. After the broader futuristic technology section, COMMS participants were asked two questions specifically about NG 911:

1. "Have you ever heard of Next Generation 9-1-1?"
2. "Next Generation 9-1-1 is a system that will allow the public to send texts, pictures, and video to 9-1-1 call centers. Do you think this will help you in your job?"

The response items for these questions were: Yes, No, or Not Sure. Interview data drove the design of the survey and indicated that some first responders did not know what NG 911 was or how it would apply to their work. The survey intentionally used a simplified definition of NG 911 in the second question listed above; content expert reviewers of the survey believed it better captured how COMMS participants would define and understand it.

Virtual Reality Question. Given its importance to PSCR's initial research agenda [15], all participants were asked specific questions about the use of VR for training and for other purposes. The two questions asked were:

1. "Do you think VR (virtual reality) would be useful for training in your work?"
2. "Do you see VR as useful in other ways for your work?"

The response items for these questions were: Yes, No, or Not Sure. An open-ended text box was also provided to give participants the opportunity to respond with additional details about their answers to the VR questions listed above.

Survey Sampling and Dissemination. In order to reach a large number of first responders, outreach occurred at the department/agency level. The sampling frame consisted of an online database with contacts in all 10 U.S. FEMA Regions [12] and a variety of first responder departments and agencies. Other means of outreach were via public safety organizations and through previous points of contact within the public safety community. Individuals contacted were asked to forward the survey to their first responder communities and colleagues in order to reach as many departments and agencies as possible, and through them to reach first responders, in order to have broad representation. The survey was disseminated to first responders across the U.S. for approximately 5 months between 2018 and 2019.

3 Participants

The first responder population sample for the interviews and survey accounted for geographic and cultural diversity; different area types (urban, suburban, and rural); and various levels in the chain of command within the COMMS, EMS, FF, and LE disciplines. The participants in the interviews represented 13 states in eight FEMA regions; the survey had representation from all 50 states and Washington D.C. Other demographic variables of interest—such as jurisdictional level (local, county, state, federal), years of service, and age—also showed good variability in both the interview and survey data. 193 first responders participated in the interviews; 7,281 first responders completed the survey.

The 193 first responders interviewed resulted in 158 interview transcripts. Some interviews included multiple participants; five participants opted to not be recorded [4]. Each of the four disciplines was represented in the sample; **Fig. 1** below shows a breakdown of interview participants by discipline and area type.

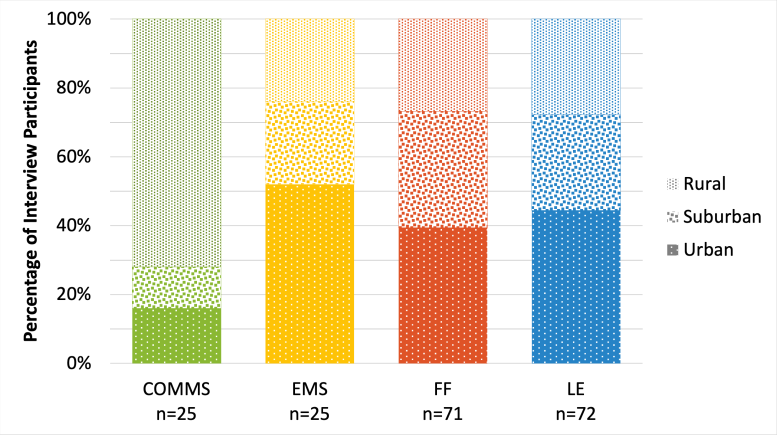


Fig. 1. Interview participants by area type

Likewise, the survey sample included diverse representation of the first responder population in all four disciplines. **Fig. 2** shows a similar breakdown for the survey data – participants who completed the survey by discipline and area type.

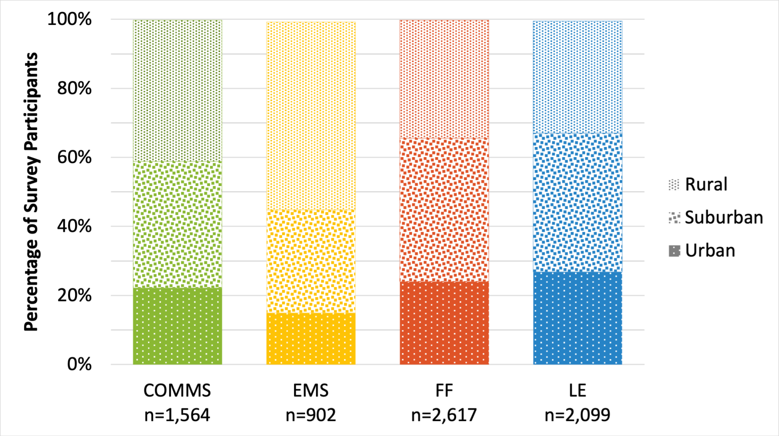


Fig. 2. Survey participants by area type

4 Results

The results presented here are quantitative survey data supported by qualitative data from both the survey (from open-ended survey questions) and interviews. Quotes from the qualitative data are verbatim and are indented in blue text with a reference notation following each quote. The reference notation represents a particular participant response and is composed of three parts: the first represents the discipline of the response (COMMS; EMS; FF; LE), the second represents the area of the response (Urban=U; Suburban=S; Rural=R), and the third is the record ID number. Interview quotes are distinguishable from survey quotes in their notations; "INT" precedes the three-part notation for interview quotes. For example, (FF:U:1234) represents the survey responses for record ID #1234, from a fire service respondent in an urban area; (INT-LE-U-006) refers to an LE interview, from an urban location, who is law enforcement interviewee number 006. It is important to highlight that these notations are not connected to specific participants, as survey and interview data are anonymous.

As previously stated, this paper focuses on a subset of the survey data that are the results of the analysis of three survey questions: 1) futuristic technologies; 2) NG 911 (COMMS only), and 3) VR. Due to the relationships between the responses to these questions and the complexities in the data, the presentation of the results is structured as follows. Examined first is the usefulness of existing devices to which first responders do not have access (deeming them "futuristic"), both across and within the four disciplines (see Sect. 4.1). Second, technologies that are typically considered futuristic both within the public safety domain and externally (e.g., VR) are explored across disciplines (see Sect. 4.2). Finally, the paper presents the discipline-specific communication technologies that first responders think would be most useful for incident response, including, for COMMS, NG 911 (see Sect. 4.3).

The analyses yielding these results was performed on unweighted survey data (see Appendix). Survey responses are representative of the first responders who completed the survey; weighting of the data should be applied prior to making any generalizations about the results to the broader public safety population. The full dataset from the interviews and survey are available online [16].

4.1 Access to Existing Technologies

As noted in the survey methodology section, some devices currently used for public safety communication are not used universally; while many devices currently exist for first responders use, not all first responders use or even have access to the same types of technology. Those devices that survey participants indicated that they did not have were piped forward in the survey to the list of technologies for the futuristic technology question. As expected, these devices varied across disciplines and demographic measures, including technologies that are often considered to be more mainstream in the public safety domain today. Perhaps of most importance here are all the basic items that respondents still do not have, but that they believe would be useful, e.g., radios and mobile data terminals (MDTs).

Across the four disciplines, survey respondents consistently identified work-issued smartphones as something that would be useful in their day-to-day work; 21.08% of COMMS thought they would be useful, 31.11% of EMS, 30.41% of FF, and 39.34% of LE. While smartphone technology exists, many first responders do not currently have access to work-issued smartphones. In contrast to the usefulness of work-issued smartphones, much lower percentages of participants thought personal smartphones would be beneficial; 8.96% of COMMS, 13.40% of EMS, 19.34% of FF, 8.05% of LE. Dawkins, et. al. posited that the concerns over the cost of smartphones could explain the discrepancy between work-issued and personal smartphones in the perceived benefits of their use [11]. The interview data mirrored these findings, showing the lack of access is often due to the cost of the devices as well as the additional costs beyond the technology itself, such as maintenance and data plans.

> At this point, I would love to buy officers smart phones, but I don't have the funding for it. So right now the only communication device that the department supplies is the radio. (INT-LE-U-029)

In addition to cost, particularly for personal smartphones, research findings suggest that major detractors from smartphones' usefulness to first responders were due to the necessity for personal data plans, the lack of adequate (if any) subsidies, and the possibilities for the subpoena of a first responder's personal smartphone [11].

Aside from smartphones, the only other technology existing in the public safety domain that crossed all four disciplines in a similar manner was desktop computers. Desktop computers are not typically considered a futuristic technology, yet like work-issued smartphones, are a technology that not all first responders currently have access to in their day-to-day work. While the percentages of EMS and FF who chose this item were somewhat low (EMS—11.65%, FF—9.58%), far more LE and COMMS respondents chose this item (19.39% of LE and 38.46% of COMMS).

Several other technologies were included in the list for three of the four first responder disciplines—EMS, FF, and LE—those public safety disciplines that are in the field. At least 20% of participants in each of these three disciplines thought the following devices that they do not currently have would be useful for their work:

- Laptop computer
- Mobile Data Terminal (MDT) or Mobile Data Computers (MDCs)
- Portable radio
- Tablet
- Vehicle radio
- Work-issued wireless earpiece

These devices, particularly MDTs and radios, represent critical public safety communication devices – something identified in the interviews as very important to first responders [4]. Again, these do not represent new or especially futuristic technology, but they are items that many first responders do not currently have but identify as potentially useful for their day-to-day incident response.

In addition to these cross-cutting technologies are those discipline-specific technologies to which first responders do not currently have access. These discipline-specific items were often those chosen by the largest percentage of respondents within the disciplines who use them. Fingerprint scanners (45.59%) and license plate readers (46.11%) were the top two devices chosen by LE respondents, with body cameras also chosen by a large percentage (31.96%). Thermal imaging cameras (TIC) for FF (27.15%) and headsets (32.47%) for COMMS also represent discipline-specific items that were selected by large percentages of their corresponding respondents. While these represent discipline-specific needs, large percentages of respondents who did not have access to them identified them as useful for their day-to-day work.

As public safety looks toward the use of more cutting-edge technologies, it is important to consider ways to make the existing technologies presented to this point more accessible to first responders in order to appropriately address the needs of the public safety community.

4.2 Technologies Useful for All

The majority of the technologies listed for the futuristic survey question were predetermined during survey development (see Sect. 2.3). Several of the technologies listed for all four disciplines – COMMS, EMS, FF, and LE – were selected by high percentages of respondents. The one item that over 50% of respondents in each discipline chose was "one login" (instead of many different usernames and passwords). While not yet ubiquitous, the use of one login, or single sign-on (SSO), is becoming increasingly widespread for the general public, but is still uncommon in public safety—for first responders, one login is still "futuristic" technology. One login was the top overall item checked for FF and LE, and the second overall item for COMMS and EMS, demonstrating its importance across all four disciplines (see **Fig. 3**). This mirrors the findings from the interview data – a major source of frustration for many first responders was the requirement to use multiple logins and passwords on their devices [4, 11]. The open-ended survey responses also indicate that SSO would be of tremendous benefit for first responders.

> One login would be at the top of everybody's list here. It is ridiculous the number of passwords and log-ins that have to be used and waste the time of first responders in their preparation and continuous log-in status. (LE:R:5075)
>
> I need to purchase an app just to remember all of the id's and passwords I need for each program I need to use. This is very frustrating and time-consuming. Where is the fob that allows me to log into anything I want? Biometrics? Bring it! (FF:S:4460)
>
> ONE LOGIN!!! Gosh, I spend an inordinate amount of brainspace and time tracking all my logins. (COMMS:U:3213)

These open-ended survey responses highlight the quantitative survey data about the importance of having one login, showing that first responders believe SSO would save time and lead to less frustration.

Three other technologies garnered relatively high percentages from first responders in all four disciplines, making them the desired future, in part, of the broader public safety domain: real-time on-scene video, indoor mapping, and voice controls for hands-free input (see **Fig. 3**).

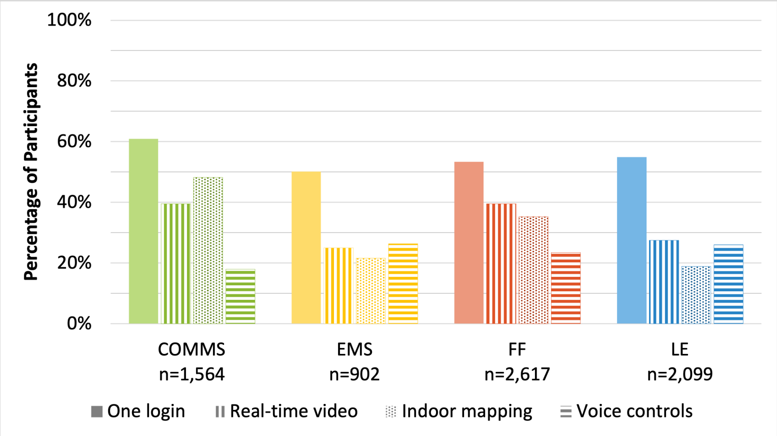


Fig. 3. Top futuristic technologies across all four disciplines

While these technologies were identified by all four disciplines as potentially useful for their day-to-day work, there were some differences amongst the disciplines. For example, COMMS and FF respondents chose indoor mapping and real-time on-scene video more often than their EMS and LE colleagues, while there was greater consistency across disciplines for voice controls for hands-free input.

As with the data presented in the previous section, some technologies cut across the three disciplines for which first responders work in the field —EMS, FF, and LE. When asked if drones would be beneficial in their day-to-day work, large percentages of FF and LE thought they would, while fewer EMS thought drones would be beneficial (see **Fig. 4**). However, in each of these three disciplines, including EMS, drones were one of the technologies that intrigued first responders during the interviews. First responders expressed how both aerial drones (e.g., to give "a live feed 360 view of [the scene]" (INT- FF-S-033)) and ground drones (e.g., "the BB-8 character from Star Wars… get that little ball with the camera… [for] reconnaissance." (INT-LE-U-013)) would be useful for incident response [7].

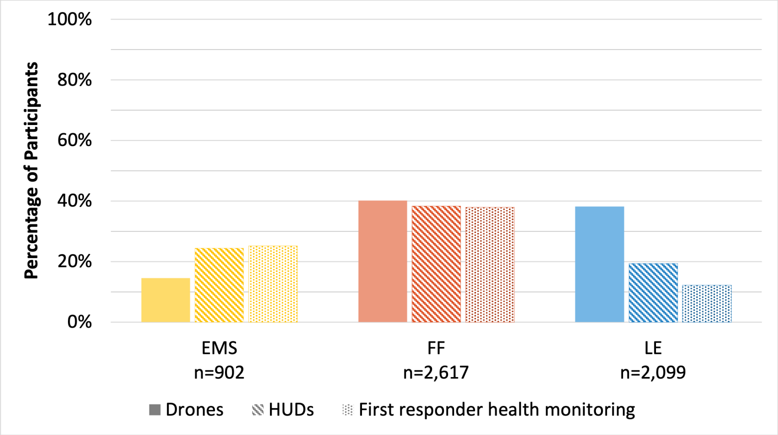


Fig. 4. Top futuristic technologies across EMS, FF, and LE

The highest percentage of participants who thought heads-up displays (HUDs) would be beneficial were in FF. Interview data show that FF envision HUDs built-into their face pieces, where they "can glance down at that HUD and look through the thermal imager if the smoke is too thick [to] be able to see through otherwise" (INT-FF-S-040). The status of first responder health and vitals is also a critical piece of information in high-risk environments; many FF and EMS respondents thought the health monitoring of first responders would be beneficial to their work. EMS and FF first responders' priority is preservation of life, but this information would also be especially helpful for incident commanders managing an incident.

Finally, the survey asked participants about the use of VR in multiple ways. First, VR was included in the list of technologies for the futuristic question. Second, the survey asked if VR would be useful for training. Lastly, participants were asked about the potential use of VR for other purposes. Results show that the usefulness of VR to first responders was tied to the way it would be used in their work contexts.

When asked about general VR benefits in their day-to-day work and about other uses for VR, respondents either did not think it would be helpful or were unsure about its usefulness. A very low percentage of respondents selected VR in the list of futuristic technologies; less than 7% in each discipline thought VR would be useful in their day-to-day work (4.92% COMMS, 3.33% EMS, 6.84% FF, 5.81% LE)[1]. In comparison with the other futuristic technologies listed, these data suggest that there are far more technologies that first responders think would be useful in their day-to-day work than VR (see Appendix). This is demonstrated further in the results of the question asking participants if VR would be useful in their work for purposes other than training. Over 50% of respondents in each discipline responded, “Not sure,” indicating that first responders were unsure if VR would be useful in other ways for their work. In fact, more respondents in all four disciplines chose “No” than “Yes” in response to this question.

While respondents had difficulty imagining other situations in which VR might be useful, when asked to think specifically about VR and training, they were more able to recognize its potential utility. Responses on the use of VR for training show more than 50% of respondents from EMS (50.28%), FF (51.54%), and LE (58.83%) said they believe VR would be useful for training in their discipline (see **Fig. 5**). For COMMS, this percentage was slightly lower at 33.78%, but still higher than the percentage of COMMS respondents who indicated they did not see VR as useful for training in their work. While high numbers of respondents supported the use of VR for training in their discipline, it must be noted as well that over 30 % of respondents from all four disciplines indicated they were not sure if VR would be useful for training in their work. These data show that many first responders need additional information about the potential capabilities and value of VR to their work if VR is to be used in public safety.

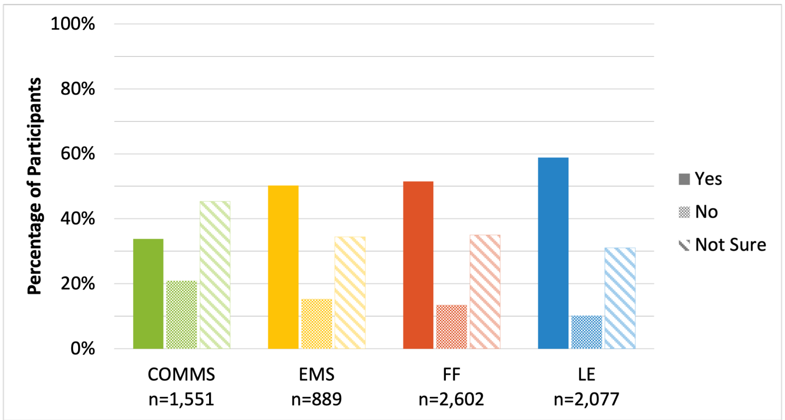


Fig. 5. Usefulness of VR for training

[1] Similarly, low percentages of respondents thought augmented reality (AR) would be useful in their day-to-day work: 4.80% of COMMS, 4.55% of EMS, 5.88% of FF, 4.95% of LE.

The higher percentage of LE respondents who saw VR as useful for training may be due to their familiarity with simulation-based training in general, while the COMMS percentage may be lower since their work is often based on audio rather than video or in-person interaction. While some first responders see the benefits of VR for training, there are others who feel VR would be a hinderance to the work of first responders, especially in an operations capacity.

> Training, but I am not yet able to see the applicability of VR in the day-to-day operations. (EMS:S:2482)
> I do not see its practical application. (FF:S:250)
> So far, everything I've seen about VR seems gimmicky - more of a toy than a useful technology. (LE:R:4511)
> VR, to me, seems to be a system for gaming and entertainment… I unfortunately see little practical application it could be used for in 9-1-1 dispatch at this time. (COMMS:S:46)

Overall, the quantitative survey data related to VR show support across all four first responder disciplines for the use of VR for training in public safety. However, as these quotes show, the open-ended data are somewhat more qualified, with first responders noting other factors that affect VR's utility, even for training.

4.3 Discipline-Specific Technologies

For the futuristic survey question, some of the technologies listed were discipline-specific due to the various types of needs of first responders (see Sect. 2.3). These technologies provide specific functions and support for first responders and are of tremendous importance to the disciplines that use them. The subsequent sections here are centered around a single discipline in presentation of these data.

COMMS. First responders in Comm Center & 9-1-1 Services have unique roles in unique environments within public safety. As such, the communication technology used in COMMS is quite different than the other disciplines, which is reflected in the survey design as well as the results. 71.23% of COMMS respondents thought automatic caller location would be useful in their day-to-day work, far more than the other futuristic technologies listed for the futuristic survey question. A key component of the day-to-day work in COMMS is interacting with 9-1-1 callers and relaying their information to first responders in the field. With the ever-increasing number of 9-1-1 calls from mobile devices, accurate location of callers is essential to their work. Another technology that a high percentage of COMMS thought would be useful is first responder tracking; 60.55% of respondents selected this technology. As COMMS represents both call taking and dispatching responsibilities, first responder tracking would have a major impact on the day-to-day work of COMMS dispatchers.

As discussed in Sect. 2.3, the COMMS survey was uniquely positioned to include questions about NG 911. **Fig. 6** depicts results showing that COMMS respondents overwhelmingly said they had heard of NG 911 (89.72%) and believed it will be helpful in their work (74.47%). The fact that almost 20% of respondents (19.55%) said they were not sure that NG 911 will be helpful in their work may demonstrate a lack of clarity about NG 911 and the ways in which it might benefit COMMS workers.

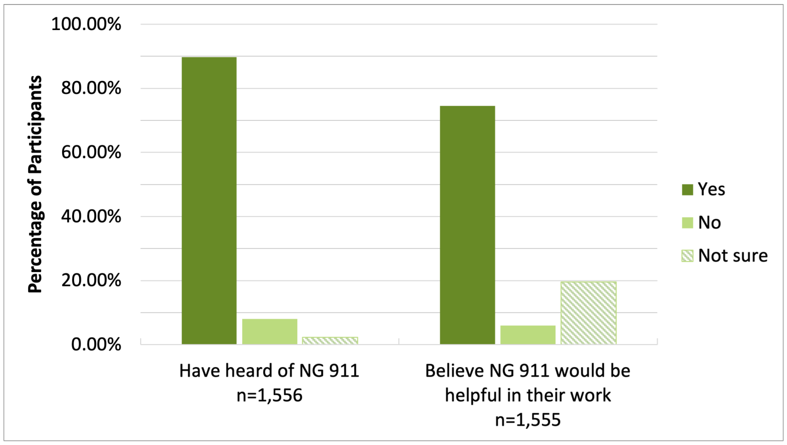


Fig. 6. COMMS survey responses to NG 911 questions

EMS. For EMS, more than half of respondents thought automatic transmission of patient vitals and information to the hospital would be useful in their day-to-day work (56.43%). Nearly 40% also thought health/vitals monitoring of patients and automatic vehicle location (AVL) would be useful (39.47% and 39.36%, respectively). The primary task for day-to-day work in EMS is treating patients. It is understandable that improvements to the health monitoring of EMS patients, as well as automating their communication tasks while treating those patients, are desirable technologies for the future of EMS.

FF. The fire service is unique in that many of FF first responders are cross-trained in EMS – their responsibilities include both fire-related and health-related service. As a result, there is some overlap in the future of communication technology with EMS and FF. This is reflected in the discipline-specific survey results, where nearly half of FF thought AVL – technology that enables COMMS to dispatch the closest vehicle to an incident, rather than just the closest station – would be useful in their day-to-day work (49.41%); a high percentage of EMS also thought AVL would be useful.

LE. Two discipline-specific technologies were selected by nearly 40% of LE – facial recognition and thermal imaging. In their day-to-day work, first responders in LE regularly need to identify persons of interest. Results suggest that first responders think technology may help in this task, as 38.69% of LE respondents thought facial recognition software would be useful in their day-to-day work.

As previously stated, some technologies listed for the futuristic survey question are more commonly used by the general public, but not as widely used in public safety; other technologies are used by some agencies and departments in public safety, but their use is not universal. Thermal imaging falls into the latter category of technology – TICs are more common in FF (but still not universally used), but not as prevalent in LE. First responders in LE think it may be beneficial for this to change, with 38.40% indicating that thermal imaging would be useful in their day-to-day work.

5 Conclusion

First responders were asked about their vision of the future of communication technology for incident response. While some of the futuristic technologies used in the survey may not be considered futuristic in some arenas, these items have often not made their way into the world of public safety. One of the best examples of this is single sign-on (SSO). Across all four disciplines – COMMS, EMS, FF, and LE – over half of participants indicated that SSO for their devices would be most useful in their everyday work. While SSO is commonly used in industry, it addresses a universal pain point in public safety, where its use is less common.

Other, more futuristic technologies first responders thought would be useful include real-time on-scene video, indoor mapping, and voice controls for hands-free input. In addition to these technologies, first responders also envisioned the usefulness of futuristic technologies specific to their individual disciplines. COMMS thought automatic caller location would be the most beneficial for their work, while also recognizing the potential of NG 911 as the future of 9-1-1 technology. EMS saw technology to automatically send patient vitals to a hospital as the most potentially useful. Automatic vehicle location (AVL) was considered by FF as the futuristic technology with the most benefit. Lastly, LE thought drones, thermal imaging, and facial recognition to identify a person of interest would be equally beneficial in their day-to-day work.

While the survey results generally showed favorability towards futuristic technologies, the open-ended survey data revealed that first responders consistently emphasized that an obstacle to the use of futuristic technologies was cost [11]. In the interviews as well, many participants cited issues of cost and price as prohibitive factors related to the adoption of new forms of technology.

> …throw in the fact that most of us have inadequate funding (FF:S:5094)
>
> …the technology is there, it just costs so much. (INT-LE-U-010)
>
> Technology is very expensive. You don't just buy it and you're good. You've got to maintain it… You've got to upgrade it. (INT-EMS-R-008)

As noted by the EMS interviewee quoted above, it is not just the initial cost of technology that makes it unattainable, there are often auxiliary costs beyond the technology itself, such as associated maintenance, certification, technical support and training. Cost may be one reason that respondents did not see some of this technology as useful for their day-to-day incident response. Improving current technology and meeting current needs rather than buying into (literally and figuratively) totally new technology was an important consideration for the first responders who participated in both the interviews and the survey. The best technology in the world is not useful if those who need it cannot afford it.

Additionally, when asked about futuristic technology, first responders often cited the need to focus less on cutting edge technology, like VR, and more on basics and current technology needed by first responders rather than on new technology.

> Until rural areas have a comms infrastructure that can support BASIC communications the rest is a fantasy. (EMS:R:2434)
>
> None of [the futuristic technologies] sound particularly useful and some could be disruptive to our normal work processes in dispatch. (COMMS:S:1545)
>
> Instead of introducing all this extra new stuff let's, one, make sure what we have actually works better. And then, two, let's not rely on it so much. (INT-FF-U-042)

If first responders are going to accept and adopt new technologies, they need to have a better understanding of how those technologies will help them accomplish their primary tasks and provide better efficiency, effectiveness, and satisfaction than what they currently use. As reported in the findings from the interview data, “New technology is exciting, and the possibilities for it are endless. While new technology may sound good and make sense to researchers and developers, adoption requires buy-in from first responders” [4].

As technology for the NPSBN is being developed, researchers, designers, and developers alike need to focus on the needs of the users – the first responders. As we learned in our interviews, there is no room in public safety to develop “technology for technology’s sake” [4]. The interviews and survey both suggest that “one size does not fit all” – first responders are open to new and exciting technologies, but their needs are utility driven; to have the biggest impact, their communication technology must be tailored to each discipline’s needs and contexts.

Acknowledgements

NIST would like to thank the many first responders, public safety personnel, and public safety organizations who graciously gave their time and input for this project

Appendix

Table 1 and **Table 2** show the results from the responses to the survey question on futuristic technology. The number of respondents, n, for each technology and the corresponding discipline is the following, unless otherwise noted: COMMS, n=1,564; EMS, n=902; FF, n=2,617; and LE, n=2,099.

Table 1. Participants who selected preset futuristic technology

Futuristic Technology	COMMS	EMS	FF	LE
AR (augmented reality)	4.80%	4.55%	5.88%	4.95%

Futuristic Technology	COMMS	EMS	FF	LE
Automatic caller location	71.23%			
Automatic transmission of patient vitals and information to hospital		56.43%		
AVL (automatic vehicle location)		39.36%	49.41%	
Drones		14.52%	40.20%	38.21%
Facial recognition software	16.05%			38.69%
First responder tracking	60.55%			21.30%
Health/vitals monitoring of first responders		25.17%	37.87%	12.15%
Health/vitals monitoring of patients		39.47%	19.07%	
HUDs (heads-up displays)		24.39%	38.29%	19.39%
Indoor mapping	48.15%	21.51%	35.27%	18.87%
One login (instead of many different usernames and passwords)	60.93%	50.11%	53.31%	54.88%
Real-time on-scene video	39.51%	24.94%	39.47%	27.49%
Remote sensing (by aircraft or satellite)			10.58%	
Robots		2.00%	4.93%	7.86%
Self driving cars		6.21%	3.97%	3.53%
Smart buildings		6.98%	12.99%	7.58%
Smart glasses		8.98%	8.06%	7.86%
Smart watch	7.23%	16.41%	12.95%	15.39%
Thermal imaging				38.40%
Vehicle tracking				26.39%
Voice controls for hands-free input	17.90%	26.27%	23.42%	25.96%
Voice recognition for identification		15.52%	13.11%	16.77%
VR (virtual reality)	4.92%	3.33%	6.84%	5.81%

Table 2. Participants who selected existing technology

Existing Technology	COMMS	EMS	FF	LE
Body camera				31.96%; n=1214
Computer: desktop	38.46%; n=26	11.65%; n=103	9.58%; n=240	19.39%; n=196
Computer: laptop		31.62%; n=136	35.93%; n=501	38.66%; n=476
Dash camera				25.43%; n=1266
Earpiece: wireless (self purchased)		3.03%; n=661	9.32%; n=1931	4.67%; n=1799

Existing Technology	COMMS	EMS	FF	LE
Earpiece: wireless (work issued)		21.67%; n=812	28.95%; n=2297	34.74%; n=1802
Earpiece: with cord		4.47%; n=694	4.81%; n=1890	6.63%; n=1147
Fingerprint scanner				45.59%; n=1349
Flip phone: work issued		2.66%; n=788	1.77%; n=2369	2.47%; n=1906
Foot pedal	10.87%; n=276			
Headset	32.47%; n=231			
License plate reader				46.11%; n=1644
MDT/MDC (mobile data terminal/computer)		32.95%; n=516	38.98%; n=1116	28.46%; n=615
Microphone: desktop	7.16%; n=433			
Microphone: handheld or clip-on	9.08%; n=859			
Mic: wireless		15.04%; n=791	19.50%; n=2251	19.40%; n=1696
Mic: with cord		4.33%; n=393	3.46%; n=752	3.08%; n=746
Monitor (at your personal workstation)	25.00%; n=56			
Monitor (for shared viewing)	20.46%; n=391			
Pager	1.33%; n=1125	6.27%; n=383	2.97%; n=1178	0.55%; n=2002
Phone: landline	16.67%; n=42			
Radio	11.94%; n=67			
Radio: in-vehicle		21.62%; n=111	35.68%; n=213	24.38%; n=320
Radio: portable		32.79%; n=61	34.88%; n=43	11.83%; n=93
Smartphone: personal	8.96%; n=201	13.40%; n=97	19.34%; n=331	8.05%; n=410
Smartphone: work issued	21.08%; n=887	31.11%; n=601	30.41%; n=1391	39.34%; n=816
Tablet		36.50%; n=326	33.88%; n=856	23.33%; n=1380
TIC (thermal imaging camera)			27.15%; n=291	

References

1. Public Safety Communications Research: Research Portfolios. https://www.nist.gov/ctl/pscr/research-portfolios, last accessed 2021/1/11.
2. Public Safety Communications Research: User Interface/User Experience Portfolio. https://www.nist.gov/ctl/pscr/research-portfolios/user-interfaceuser-experience, last accessed 2021/1/11.
3. ISO 9241-210:2010: Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems, ISO, Geneva, Switzerland (2010) http://www.iso.org/
4. Choong, Y, Dawkins, S., Furman, S., Greene, K.K., Spickard Prettyman, S., Theofanos, M.F.: Voices of First Responders – Identifying Public Safety Communication Problems:

Findings from User-Centered Interviews. Phase 1, Volume 1. NISTIR 8216 (2018). https://doi.org/10.6028/NIST.IR.8216

5. Greene, K. K., Dawkins, S., Spickard-Prettyman S., Konkol, P., Theofanos, M. F., Mangold, K., Furman, S., Choong, Y., Steves, M. P.: Voices of First Responders—Nationwide Public Safety Communication Survey Methodology: Development, Dissemination, and Demographics. Phase 2, Volume 1. NISTIR 8288. (2020) http://doi.org/10.6028/NIST.IR.8288
6. Greene, K. K., Dawkins, S., Choong, Y., Theofanos, M. F., Prettyman, S. S., Furman, S., and Steves, M.: Characterizing First Responders' Communication Technology Needs: Towards a Standardized Usability Evaluation Methodology. In Homeland Security and Public Safety: Research, Applications and Standards, edited by Mattson, P. and Marshall, J. (West Conshohocken, PA: ASTM International, 2019), 23-48. https://doi.org/10.1520/STP161420180048
7. Dawkins, S., Choong, Y., Theofanos, M., Greene, K., Furman, S., Steves, M., and Spickard-Prettyman, S.: Voices of First Responders – Examining Public Safety Communication Problems and Requested Functionality, Findings from User-Centered Interviews. Phase 1, Volume 2.1. NISTIR 8245 (2019). http://doi.org/10.6028/NIST.IR.8245
8. Dawkins, S., Greene, K. K., Steves, M., Theofanos, M., Choong, Y., Furman, S., and Prettyman, S. S.: Public Safety Communication User Needs: Voices of First Responders. In Proceedings of the Human Factors and Ergonomics Society's Annual Meeting, Philadelphia, PA, October 1-5, 2018.
9. Greene, K. K., Dawkins, S., Theofanos, M., Steves, M., Furman, S., Choong, Y., and Spickard-Prettyman, S.: Voices of First Responders—Examining Public Safety Communication from the Rural Perspective, Findings from User-Centered Interviews, Phase 1, Volume 3. NISTIR 8277 (2019). http://doi.org/10.6028/NIST.IR.8277
10. Steves, M., Theofanos, M. F., Choong, Y., Dawkins, S, Furman, S., Greene, K. K., Spickard Prettyman, S.: Voices of First Responders – Examining Public Safety Communication from the Perspective of 9-1-1 Call Takers and Dispatchers Findings from User-Centered Interviews, Phase 1, Volume 4. NISTIR 8295. (2020). https://doi.org/10.6028/NIST.IR.8295
11. Dawkins, S., Greene, K., and Spickard-Prettyman, S.: Voices of First Responders – Nationwide Public Safety Communication Survey Findings, Mobile Devices, Applications, and Futuristic Technology. Phase 2, Volume 2. NISTIR 8314 (2020). http://doi.org/10.6028/NIST.IR.8314
12. Federal Emergency Management Agency (FEMA): Regions (2020). https://www.fema.gov/about/organization/regions, last accessed 2021/1/11.
13. Joint Program Office for Intelligent Transportation Systems: Next Generation 9-1-1 (NG9-1-1) System Initiative Concept of Operations. U.S. Department of Transportation. (April 2007). https://rosap.ntl.bts.gov/view/dot/4025, last accessed 2021/1/11.
14. National 911 Program. Next Generation 911. Office of Emergency Medical Services, National Highway Traffic Safety Administration, U.S. Department of Transportation. https://www.911.gov/issue_nextgeneration911.html, last accessed 2021/1/11.
15. Feldman, H., Leh, M., Felts, R., and Benson, J.: Public Safety User Interface R&D Roadmap. NIST Technical Note 1961. (2017). https://doi.org/10.6028/NIST.TN.1961
16. NIST PSCR Usability Team: PSCR Usability Results Tool: Voices of First Responders. NIST (Revised 2020). https://publicsafety.nist.gov/, last accessed 2021/1/11.

Current Problems, Future Needs: Voices of First Responders about Communication Technology

Kerrianne Morrison[0000-0002-2735-8809] (✉), Shanee Dawkins[0000-0002-8114-0608] (✉), Yee-Yin Choong[0000-0002-3889-6047] (✉), Mary F. Theofanos, Kristen Greene[0000-0001-7034-3672], and Susanne Furman

National Institute of Standards and Technology, Gaithersburg, MD 20899 USA
kerrianne.morrison@nist.gov, shanee.dawkins@nist.gov, yee-yin.choong@nist.gov, mary.theofanos@nist.gov, kristen.greene@nist.gov, susanne.furman@nist.gov

Abstract. With advances in network technologies, there has been increasing interest in developing new communication technology for first responders that utilizes wireless broadband networks. In order to develop new communication technology, user requirements are needed to ensure new technology is usable; however, capturing user requirements for first responders has been challenging due to the diversity of its users' contexts and needs. This paper aims to provide guidance and insight into developing user requirements for communication technology developed for first responders by exploring first responders' communication technology problems and needs. Qualitative interviews with 193 first responders across four disciplines (Communications (Comm) Center & 9-1-1 Services; Emergency Medical Services; Fire Service; Law Enforcement) revealed that they often encountered problems with their communication technology's reliability, usability, and interoperability. Their primary need for their communication technology was for solutions to their current problems, rather than development of new technology. Many were also interested in communication technology that can provide them with real-time information. This study underscores that communication technology for first responders should be designed and developed for and with first responders.

Keywords: Cognition · First Responders · Psychology · Public Safety · Social Sciences · Usability · User Requirements · UX

1 Introduction

First responders play a critical role in maintaining public safety, as they are the frontline responders in emergencies and major crises. During these incidents, first responders rely on communication technology such as radios, cell phones, and computer-aided dispatch (CAD) to gather information and coordinate appropriate incident response.

Buchanan, Kerrianne; Dawkins, Shanee; Choong, Yee-Yin; Theofanos, Mary; Greene, Kristen K.; Furman, Susanne M.; Prettyman, Sandra. "Current Problems, Future Needs: Voices of First Responders about Communication Technology." Presented at 23rd International Conference on Human-Computer Interaction (HCI International 2021). July 24, 2021 - July 29, 2021.

Historically, first responders have used land mobile radios (LMR) to communicate and share information during incident response [1]. Although first responders have primarily relied upon LMR technology for their incident response, first responders have increasingly begun to utilize technology compatible with wireless broadband network solutions to communicate and transmit information [2].

Until recently, there was no dedicated wireless broadband network infrastructure for first responders in the United States (U.S.). Fortunately, in 2012, the Middle Class Tax Relief and Job Creation Act provided funding to support the creation of the Nationwide Public Safety Broadband Network (NPSBN) [3]. This system, currently being deployed, will allow for long-term evolution (LTE) data-based solutions to supplement LMR. Utilizing data-based technology has the potential to improve the efficiency and effectiveness with which first responders respond to incidents, as research suggests that new devices connected through the broadband network may allow them to transmit and access data and information more quickly ([4-6]) and access new types of information and functions previously not available to them (e.g., Internet of Things devices [7]).

Although the technological benefits first responders may derive from communication technology and other types of network-enabled technology are clear, the user requirements needed for this new technology have yet to be determined. Determining appropriate user requirements is a challenging task due to the specific and unique needs of the first responder user population. First responders require communication technology and equipment that will perform specific functions (e.g., push-to-talk) in a variety of environments (e.g., fires, rural areas) and scenarios (e.g., emergency crisis, major disasters) [2,8]. An added challenge to developing user requirements for first responders is the diversity of disciplines within the first responder population. Disciplines such as Communications (Comm) Center & 9-1-1 Services (COMMS), Emergency Medical Services (EMS), Fire Service (FF), and Law Enforcement (LE) have different needs due to the variety of duties they perform and contexts in which they work [2,8]. First responders also have unique infrastructure, functionality, and economic (e.g., market size and budgets) needs for technology. These needs differ greatly from other populations such as the general public or even the military population, who often respond to similar scenarios (e.g., major disasters) [2]. Therefore, the distinctive needs of the first responder population often preclude developers from simply retrofitting user requirements and technology from other populations for first responders' use [2]. Research and development are needed to develop unique and specific user requirements and technology for first responders that account for the contexts they work within and the needs they have.

To develop these user requirements, a user-centered design approach is needed. This approach posits that to develop technology, it is important that the characteristics, experiences and needs of the users of technology are taken into account, considering their environments, context of work, and needs when developing technology for them [9,10]. Ultimately, taking users into account helps to ensure new technology is usable, enabling users to complete their desired tasks with effectiveness, efficiency, and satisfaction [9,10]. Although some work has used user-centered design approaches when developing and designing technology for first responders (e.g., [4]), less work to date has comprehensively examined the context and needs of the first responder population.

The National Institute of Standards and Technology (NIST) Public Safety Communications Research (PSCR) program focuses on improving first responders' communication technology by conducting research and development efforts (see [11]), with User Interfaces and User Experiences (UI/UX) as one of the key areas for research and development [12]. To support UI/UX research and development, the NIST PSCR Usability Team utilizes human factors and user-centered design principles to produce guidance and insight for gathering user requirements for researchers, developers, and designers [13]. As part of this effort, an exploratory, sequential, mixed-methods study was conducted with first responders. In this multi-phase study, an initial qualitative phase (Phase 1) was followed by a quantitative phase (Phase 2) to investigate first responders' experiences with and perceptions of communication technology [8,14]. The research questions for this project were:

1. How do public safety personnel describe the context of their work, including their roles and responsibilities as well as process and flow?
2. How do public safety personnel describe their communication and technology needs related to work?
3. What do public safety personnel believe is working or not working in their current operational environment related to communication and technology?

In Phase 1 of the project, interviews were conducted with first responders to understand their context of work and their communication technology experiences, including their most pressing problems and needs [8]. Results from initial analyses of the Phase 1 interviews suggested that first responders' needs are specific to their environments and tasks, and that a "one size fits all" approach to developing communication technology is ill-suited to improving first responders' communication technology [8,15].

Phase 1 data analysis also included gaining an in-depth understanding of what first responders' current problems are with their communication technology as well as what their top requested functionalities are for how they want their communication technology to work [16]. This paper presents the results of this analysis, examining problems and requested functionalities to highlight the unique needs of the first responders interviewed across disciplines. Ultimately, results from this paper will provide insights and guidance for developers and designers to create and improve communication technology tailored to the needs of first responders.

2 Methods

As previously mentioned, a multi-phase mixed methods study was conducted to examine human factors issues for first responders with their communication technology. In Phase 1, qualitative interviews were conducted [8,15]. Findings from Phase 1 then informed the design of the survey used in Phase 2 [14,17].

This paper focuses on the qualitative data obtained as part of Phase 1. Conducting a qualitative study had many advantages. First, a qualitative approach allowed for an in-depth and contextualized understanding of first responders and their perspectives on their work, environment, and communication technology. Second, using semi-structured interviews allowed for interviews to be dynamic, exploring the topics and needs that were top of mind for first responders. Third, this approach allowed for inclusion of first responders in the research process, aligning with user-centered design principles [9,10].

2.1 Data Collection

Recruitment. To ensure the study represented the breadth of first responders, purposive sampling was used to recruit a sample that varied across first responder disciplines in urban, suburban, and rural areas of the U.S. First responders across four disciplines were represented in the interviews: COMMS, EMS, FF, and LE. First responders of different ages and genders as well as various levels of experience also participated. This resulted in a sample of 193 first responders from across the U.S.

Procedure. Interviews with first responders occurred at their place of work in a private room or area. A primary goal of the research design was to reduce the burden as much as possible on first responders; therefore, the interview session was designed to last approximately 45 minutes. Moreover, to increase efficiency, some interviews were conducted with multiple first responders at a time. This resulted in 158 interviews with 193 first responders. With permission, the research team recorded interviews so that they could be transcribed later for use in the data analysis process. In five interviews, participants opted out of recording; in these instances, the interviewer's notes served as interview data. Prior to the interview, participants completed a short demographic form.

All participants were informed that their participation was voluntary and that they could leave the interview at any time. The NIST Research Protections Office reviewed the protocol for this project and determined it met the criteria for "exempt human subjects research" as defined in 15 CFR 27, the Common Rule for the Protection of Human Subjects.

2.2 Measures

Interview Methodology. A semi-structured interview protocol was developed for interviews based on the project's research questions and a review of prior literature. The instrument also was reviewed by subject matter experts and piloted by first responders. This helped ensure the final instrument's content sufficiently addressed the research questions. It also served as a check that the content and language were appropriate for first responders. These reviews determined that two interview protocols were needed: one for EMS, FF, and LE who respond to incidents on-scene, and one for COMMS who coordinate incident response off-scene in communications centers.

The instrument included questions about first responders' context of work and communication technology. The context of work questions focused on the first responders' roles, daily tasks and routines, interactions with other people, and specific work environments. Questions about first responders' perceptions of and experiences with communication technology focused on what kinds of technology and information they use and what problems they encounter with communication technology. The instrument also included questions that asked them to describe their top requested functionalities: if they could have anything, what communication technology they would want? This paper focuses on responses to queries about communication technology problems and their requested functionalities.

Demographics Questionnaire. The demographics questionnaire asked for participants' genders, ages, years of service, and technology experience and adoption. These questions were also assessed by subject matter experts and piloted by first responders.

2.3 Demographics

All four first responder disciplines were represented in the data, with most participation from first responders in FF (n=71; 36.8%) and LE (n=72; 37.3%) disciplines and fewer from COMMS (n=25; 13.0%) and EMS (n=25; 13.0%). Fig. 1 displays the percentage of participants in each discipline by area.

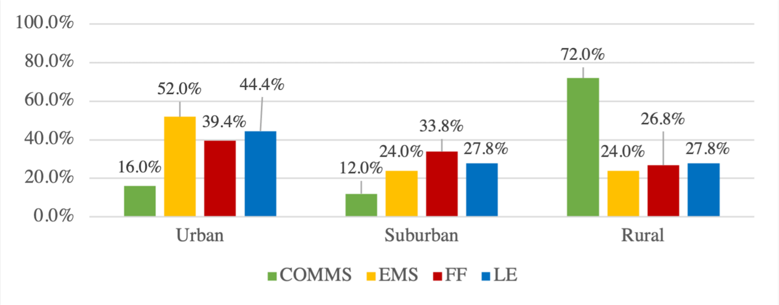


Fig. 1. Participants' disciplines by area

First responders from across urban (39.90%), suburban (27.46%), and rural (32.64%) areas were interviewed in eight of the ten U.S. Federal Emergency Management Agency (FEMA) regions [18]. A majority (86%) of the sample was male, which is consistent with the rates found nationwide for FF [19] and LE [20] first responder populations. Fig. 2 displays the participants' ages and years of service (note that two participants did not answer these demographic questions). A majority of participants were between 26 and 55 years old, and experience in public safety ranged from one year to over 30 years of experience.

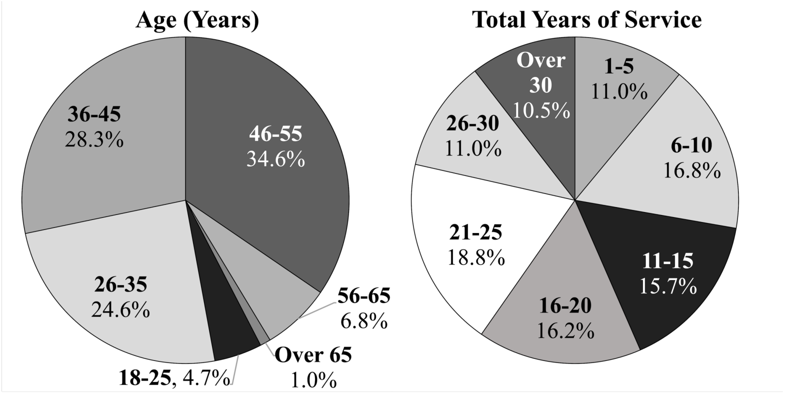


Fig. 2. Demographic characteristics

2.4 Data Analysis

To analyze the interview data as part of the larger project, transcripts were coded. In qualitative research, coding is the process of labeling or categorizing participant responses so that they can be extracted to identify themes in the data. For this project, the coding process first started by generating an *a priori* list of coding categories based on the research questions and literature review. The research team then used the coding categories to code all responses in five transcripts. The researchers met to discuss how each coding category was defined and applied to responses. This discussion resulted in the formation of a final coding list that was then applied in a consistent manner to the remaining transcripts. Once transcripts were coded, the coded responses were extracted together into a document where researchers identified themes across responses. Choong et al. [8] describes the coding process and resulting categories and themes in detail.

To provide a more in-depth look at first responders' problems and needs, additional analyses of the interview data were performed by further examining two of the initial coding categories: "Problems with Technology" and "Wish List". The coding category "Problems with Technology" captured responses in which problems with communication technology were discussed. The coding category "Wish List" captured responses in which first responders described their requested functionalities for how they wanted their communication technology to work. Responses for each coding category were separately extracted, meaning that data associated with each code were collected into separate documents. Once in these documents, responses were further categorized into more specific categories and subcategories to capture deeper nuances and insights.

Independent researchers examined problems and requested functionalities separately and developed the categories and subcategories. Category and subcategory lists included items that could be found in interviews across disciplines, but some were specific to certain disciplines. For instance, the problems category "9-1-1 calls" was only categorized in COMMS interviews. Additionally, because a single quote may relate to multiple categories and subcategories, there was overlap in the responses within and between categories. After an initial classification, the researchers discussed their categories and subcategories to generate a final list used to categorize all relevant responses. The analysis of the responses related to problems resulted in 1,729 quotes in 25 categories; analysis of the responses related to requested functionalities resulted in 1,143 quotes in 18 categories. Communication technology problems categories and subcategories are displayed in Table 1 and requested functionality categories and subcategories are displayed in Table 2.

Table 1. Technology Problems Categories and Subcategories

Category	**Subcategories**
9-1-1 Calls	Next Generation 9-1-1 (NG 911), caller location, nuisance calls
Audio Clarity	Hard to hear, audio feedback
Body Camera	Functional issues, physical issues
Connectivity	Reception, bandwidth issue
Disruption of Operations	Continuity of Operations (COOP), mobile operations
Implementation/IT Infrastructure	Implementation/Installation issues, cost as a prohibitor, IT management, no user requirements collected/considered, public safety network reservations
Interoperability	External interoperability, internal interoperability
Microphone/Earpiece	Cord, earpiece, wireless microphones
Mobile Data Computer (MDC)/ Mobile Data Terminal (MDT)	Navigation/mapping, functionality
Overwhelmed	Sensory overload, situational awareness
Physical Ergonomics	Robustness, battery problems, bulky and heavy, too many devices, physical discomfort, display size, safety concerns
Radio	Dead zones, traffic, channel switching, usability
Reliability	Unreliable technology, redundancy, unreliable transmissions
Security Constraints	Authentication, access control
Technology Outdated	Outdated, incomparable to personal technology
Technology Overrated	Problems with new technology, doesn't solve communication problems
User Interfaces	Ineffective and inefficient, alerting, modality
Video	Data issues, surveillance videos

Table 2. Requested Functionality Categories and Subcategories

Category	Subcategories
All-In-One	Cell phones and/or radios, tablets, software and apps, general multifunctional devices, cameras
Communications Center Technology	Improved dispatch interface, multimedia data package, access to caller cell camera, large multi-view display
Functionality	Reliability, better coverage, clearer communication, improved functionality, longer battery life, faster devices
Futuristic	Media/Science-fiction influenced, smart buildings, face and object recognition software, self-driving vehicles, augmented reality (AR), emergency traffic light system
Integrated Gear/ Wearables	Heads-up display (HUD), in-mask microphone/earpiece, responder vitals, personal protective equipment (PPE) technology
Interoperability	Software/hardware compatibility, interagency communication system, patient care reporting (PCR), body camera integration, interjurisdictional criminal data
Microphones/Earpieces	Wireless, specialized earpieces
Mobile Apps	Information references, discipline-specific apps
Physical Ergonomics	Smaller and lighter, fewer devices, robustness, larger devices
Radios	Channel switching, multiple talk groups, prevent accidental transmissions
Real-Time Technology	Live video and images - capture/live feed technology, traffic and navigation, drones, language translation, identification devices
Tracking	Responder location, caller location, search technology
Usable security	Single sign-on
User Interfaces	User friendly, hands free, non-verbal communication
Vehicles	Windshield HUD, built-in camera, automatic license plate reader, dashboard computer

3 Results

This section presents key themes found in the data for first responders' problems and needs for communication technology. Along with findings and themes, representative quotes from participants are presented. These quotes are meant to encapsulate themes from the data rather than depicting a single person's perspectives. All quotes are anonymous and cannot be tied back to participants. To provide context for the quote, each is presented with an identifier to show the discipline (COMMS, EMS, FF, LE), area (U = Urban; S = Suburban, R = Rural), and the interview number.

Cross-discipline findings are presented first, followed by discipline-specific sections to highlight problems and requested functionalities emphasized in interviews of participants from each discipline. Problems and requested functionalities are presented together as the first responders' requests were often to have their current problems addressed.

3.1 Across Discipline Results

The communication technology problems and requested functionality categories displayed in Fig. 3 and Fig. 4, respectively, are proportional to responses in the data. Although a variety of problems and requested functionalities were identified, four key areas for improvement for communication technology emerged in the data: reliability, usability (e.g., physical ergonomics, user interfaces), interoperability, and real-time technology. While problems with radios was the top problem category, this paper focuses more broadly on problems and needs that may impact multiple communication technologies.

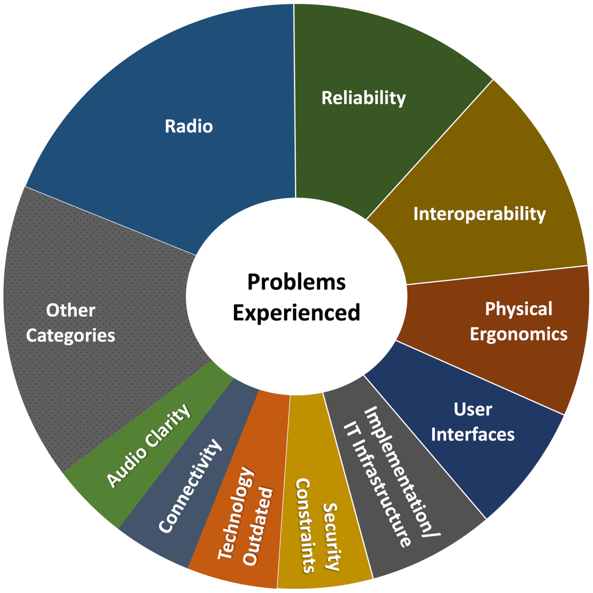


Fig. 3. Technology problems

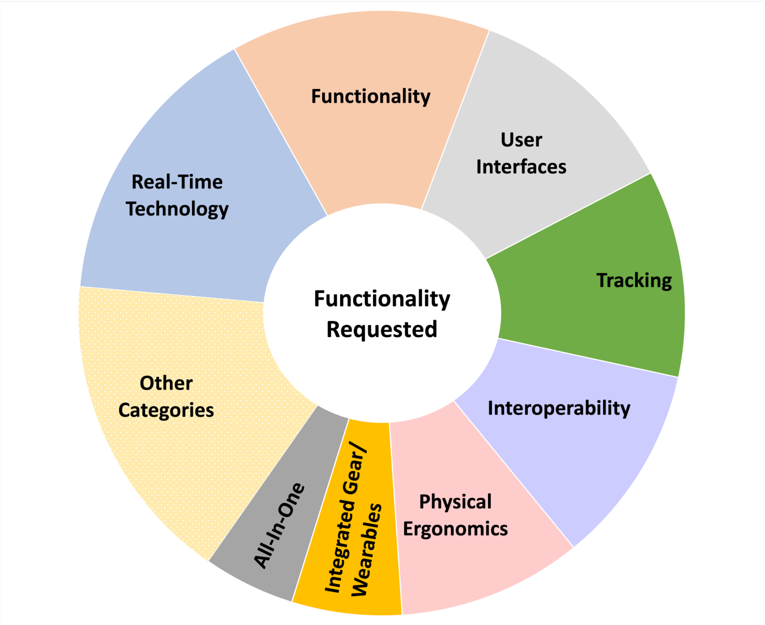


Fig. 4. Requested functionalities

Reliability. Many participants felt their communication technology was often unreliable, describing experiences in which their technology failed or did not work in the way intended. Often first responders expressed that they expected their communication technology to fail, causing them to devise back up plans and redundant systems to ensure they could still perform vital tasks during incident response.

> "In any type of critical incident, that is one of the first things that goes is the communications on cell phones… we try to have the redundant systems in place, probably more so in public safety than any other profession; we try to have those redundant systems but it still doesn't mean that, you know, some of them haven't been utilized in years and we are not sure that they are going to continue to work if we need them to do so." (LE-R-059)

Because of these failures, first responders often did not trust their communication technology. This lack of trust often resulted in first responders abandoning unreliable products altogether or opting for older solutions.

> "Everything we use…we don't have time to mess with it, or tweak it, or play with it. It has to work the first time, every

> time, or people will just to stop using it. They will just refuse to use it and go back to the old way." (EMS-U-003)

Often unreliability stemmed from challenges obtaining coverage and connectivity for communication technology. Many first responders mentioned that geographic dead zones (e.g., mountains) and structural dead zones (e.g., basements, tunnels) prevented their technology from successfully transmitting and receiving information.

> "The downside is when a firefighter goes down in a basement, and his radio doesn't see a repeater, he can't call for help, so the radio is useless. So the fix to that is to go to a direct channel. The downside of being on a direct channel is only people within a mile can hear the radio, so other people across the city or incoming to the fire don't get to hear what's going on before they get there until they get into range of the direct channel. So that's a conundrum. I guess there are some fixes. You can put a repeater on every chief's buggy that will also take that direct channel and put it into a repeater system that way everybody to hear it that way. That requires infrastructure, investment, and installation, and so forth, and money, so." (FF-U-016)

Improving connectivity and coverage was one of the top requested functionalities. By improving network access, first responders could continue to communicate and access information they need during incident response.

Usability. Participants expressed a wide variety of usability challenges with their communication technology. Many were related to physical problems they had with their communication technology, but they also experienced challenges with their devices' user interfaces.

Many participants indicated their communication technology was ill-suited to their environments and daily tasks. The combined weight and bulkiness of all the devices they carried impeded their ability to perform their jobs.

> "We have a lot of things that we carry. So there's so much on our belts… Just a little bit of weight off your belt is huge. So I have a hard time trying to fit everything on my belt as required in policy. Especially if you're some of the smaller officers, male or female just depending on your waist size, it's like you don't have enough real estate to fit all this stuff that's required. So a lot of people have dropped tasers or that kind of stuff…So I would hope that sometime the technology could help us remove some of all this hardware that we're carrying around all the time. That would be a big deal. I mean health wise for officers too." (LE-U-007)

As voiced by this LE participant, limited mobility was sometimes associated with putting first responders' health and safety at risk. For these reasons, many first responders were interested in having lightweight communication technology as well as technology that is better integrated, allowing them to carry fewer devices at a time. This could not only decrease the physical burden on first responders, but this could also allow them to have fewer distractions when responding to incidents.

Decreasing cognitive load was especially of interest to COMMS responders, who are often required to view multiple screens and use multiple devices simultaneously.

> "…with all the things that everybody wants to integrate. They want you to have apps, they want you to be able to bring in apps, they want you to be able to bring in photos and videos and texting and this and that. The more stuff we add-- the more computer screens, the more keyboards, the more mice. It just keeps adding, and it's the more burden." (COMMS-U-007)

In addition to physical usability challenges, first responders also described problems using user interfaces. Many described overly complicated user interfaces that did not function effectively and efficiently. Because first responders must often react quickly during incident response, they do not have time to interact with complex user interfaces.

> "When we need the technology, we need it to be simple. We don't need it to be complex because we don't have the time to work through complexities in anything technology because our decisions are instant…And so technology is great, but if it's complex, it kind of is counterproductive." (FF-U-025)

Lightweight devices that are easy to use and learn and that can be integrated into their current equipment and systems may address first responders' needs for more usable communication technology.

Interoperability. First responders emphasized that communication and coordination across disciplines, agencies, and jurisdictions was vital for successful incident response. Unfortunately, many first responders described challenges in effectively communicating due to poor interoperability of communication technology across different disciplines. These communication challenges often resulted in confusion and delays in incident response.

> "…communication is the key to either success or failure and you're only as good as your communication components and your knowledge of communication. So we're always lacking, in my opinion, when it comes to radio communication. There's always problems…getting on the right channel or being able to communicate with a different entity or different agency. You see that in, unfortunately, but you've got

> mass shootings and there's always a problem with cops being able to talk to firefighters. There are paramedics… And I don't know what the answer is…But I mean, there's always technology that's going to give you problems.” (FF-U-021)

Improved interoperability was one of the primary requested functionality categories in the interviews, as addressing this problem could improve coordination for incident response as well as allow for sharing data and other information.

Real-Time Information. Although a theme in many interviews was first responders' desire for solutions to the problems they currently experience with their communication technology, participants also expressed interest in having new technology that could provide them with real-time information. Many participants discussed that having information from videos and images may save time and allow first responders to be better prepared when they arrive on the scene of an incident.

> “If you could text message a picture to 911, and they could send it to us, that would help as far as we could look and say, ‘This car accident here and the people can't get out of their vehicle.’ A lot of times, the engine is dispatched by themselves. We could see that picture and say, ‘Hey, let's add on a truck,’ and achieve to that so we can extricate. Let's get that ball rolling sooner. Images would be pretty fantastic...” (EMS-S-005)

Although pictures and videos were often discussed, participants expressed interest in a variety of real-time technology including GPS navigation for traffic, language translation in real-time, and drones.

3.2 Discipline-Specific Results

All disciplines desired improvement to their communication technology's reliability, usability, and interoperability, but first responders in each discipline also expressed problems and needs unique to their individual contexts.

COMMS. COMMS responders' problems with communication technology were related to their unique context of work taking 9-1-1 calls and dispatching first responders from call centers and public safety answer points (PSAPs). COMMS responders mentioned challenges in planning to maintain continuity in the event that the communication centers' operations are interrupted. COMMS responders also saw utility for new technology to improve user interfaces, allowing them to more effectively access critical information or more intuitively navigate monitors and information.

However, many also expressed some concerns with new communication technology. Some discussed concerns about text to 9-1-1 and Next Generation 9-1-1, a digital-based 9-1-1 system [21].

> “I’m not going to have a 30-minute conversation over text whenever I could hear your voice and…and I can really hear are you okay. I can’t tell that over here…that’s my fear with the Next Generation 911 is are we going to lose that important piece of our communications with technology…But I'm hoping that the texting feature and the apps that we're using on smartphones and other devices that they don't take away that human communications during an emergency because 90% of all communication is nonverbal in nature. And hearing that voice, hearing the background noise of a particular call gives us so much more information than just the words that that caller is saying.” (COMMS-R-016)

As stated in the quote, many COMMS responders were concerned that these technologies may take away valuable information gained from voice calls.

COMMS responders also discussed situations in which their communication technology delayed their incident response. Some COMMS responders experienced delays in dispatching help to incidents when their equipment was not interoperable with the equipment of the disciplines with which they were working. Additionally, because of the abundance of cell phones in the general population, COMMS responders often have taken multiple bystanders’ calls for a single incident, causing high call volumes and delays in dispatching first responders to other incidents. Receiving nuisance calls, pocket dials, and unintended calls from smart devices also prevented them from sooner taking emergency calls.

Although COMMS responders discussed many challenges and unintended negative consequences of communication technology, they were interested in new communication technology, especially technology that could automatically locate callers. They mentioned that callers are often unable to accurately provide their location, which inhibits COMMS responders’ ability to dispatch first responders to the correct scene.

> “So location information is very important…this is a pretty big wish but one day I would love to be able to see you know accurate you know x, y, z coordinates… If we get a 911 hang up from a cell phone… and it comes back to-- you know, we don’t know where you are but you’re calling and you’re just screaming and then I have to go trace your phone and I trace your phone. And … even if they can give me an address, [if] it’s an apartment complex, I don’t know where you are. We can’t have the officers go check every apartment. We just can’t do it right? So knowing kind of an approximate would be really cool. That would be awesome. Anything where we can better direct people to where you are.” (COMMS-U-006)

Communication technology that can quickly and accurately provide callers’ exact locations to COMMS responders could help COMMS responders more quickly send other disciplines to the correct location to administer help.

EMS. EMS responders reported problems with efficiently and effectively sending important patient information to healthcare providers (e.g., hospitals) during incident response. Many discussed that these problems were often related to their communication technology's unreliability, radio dead zones, and connectivity issues.

> "I know we have issues with WiFi every once in a while. If I'm on the WiFi of the computer and then I drive to the hospital and now I'm inside the hospital writing my report and I'm on the hospital's WiFi and I go to leave, there is a space in between where I'm on neither network until I get away from the hospital. And I've had reports just get lost." (EMS-S-015)

Although EMS responders were interested in improved reliability and coverage, they also requested improved interoperability of medical equipment with external systems and organizations.

> "Maybe something along the lines that patient tracking because that's a good tool. Doesn't get used a lot though. I think something that would be an automatic download to the Red Cross. If we had victims, multiple victims from an incident…once we had their name and information, if we could tap a drop-down menu and say Red Cross or whatever, then that information would go right to their databank. And they would know where that person is…" (EMS-U-009)

Ultimately an improved and streamlined information sharing process could allow EMS responders to improve patient outcomes and help more patients.

FF. FF responders described problems communicating during incident response because of audio clarity issues. When responding to incidents, loud sounds (e.g., alarms, burning and crackling from fires, chainsaws running) often prevented FF responders from hearing others. Some FF responders also mentioned that their equipment (e.g., self-contained breathing apparatuses (SCBAs)) also contributed to audio clarity issues, often muffling voices and audio reception.

> "And depending on what's going on inside with other noises and things, that can sometimes challenge it. But every once in a while you get a garbled communication coming from somebody wearing a mask just because of placement of the radio and where they're talking." (FF-S-038)

Improving audio clarity and making communication clearer could improve communication and coordination during incident response.

FF responders were also interested in technology that could integrate communication technology with their personal protective equipment (PPE) and SCBAs. Some were interested in new technology such as heads-up displays (HUDs) or technology that could integrate microphones and earpieces directly into PPE.

> “A heads-up display in our face piece. So I've got the thermal imager attached to my face piece, so where I look, I've got a little heads-up display right in front of me. So I can either look out through my face piece and just see what's in the environment or I can glance down at that HUD and look through the thermal imager if the smoke is too thick for me to be able to see through otherwise. A HUD for the radio would not be a bad thing either. To put a display that mirrors the display on my radio so I can see what my coverage is, I can see what talk group I'm on. I can see all the stuff.” (FF-S-040)

As described by this FF participant, a HUD could improve FF responders’ access to information that may otherwise be difficult to obtain in conditions with limited visibility. It could provide them with information about the physical environment as well as assist them with coordinating with other FF responders during incident response.

LE. As previously mentioned, LE participants reported usability issues in their devices’ physical weight and bulkiness. Additionally, they discussed problems with body camera usability. Some mentioned body cameras were not well suited to their day-to-day incident response; because incidents often happen quickly, many described challenges in turning their body cameras on and off at the correct time. Relatedly, attaching body cameras and ensuring they stay attached was often difficult when incidents were physically challenging or in situations where equipment or clothing occluded recordings.

> “…The first week I had it…it was in my badge, because it was just where it was going to fit well for me on when I was wearing a jacket...The problem was, that would put it on my right side, which meant it was angling to the right, and suddenly I'm talking to people, and I tend to be [inaudible] up against them with my left side towards them, so it wasn't seeing anything, it was just pointing in the wrong direction, so I just had to learn how to move it around.” (LE-U-056)

Some LE responders described spending significant time and effort ensuring their body cameras were securely attached, working properly, and uploading videos effectively. In some cases, features of the body camera such as flashing indicator lights put officers in danger, revealing their location in dangerous situations. Thus, these problems may have downstream effects on LE responders’ efficiency, effectiveness, and safety.

In addition to improving body camera usability, some LE participants requested usable security. For example, some were interested in leveraging single sign-on (SSO) to improve authentication on devices.

> “[New technology is] over-complicated… just make it simple… When you would get in the car, you only needed like one password to get onto the computer. Now, you need like [five]. And they have to be all different, and it has to have a hashtag. It has to have this, a number. It has to have-- so me, I put them all on my phone, because I forget. I'm only almost 40, but I'm already forgetting things. But you have to know nine passwords to get on your technology.” (LE-U-013)

Like FF, many LE participants were also interested in better integration of communication technology with their equipment. Specifically, some participants requested integrating communication technology into their vehicles through technology such as windshield HUDs.

> “…Everything is visual on the screen, transparent to you, in front of you while you're driving…So the officers in itself will be able to see everything on the screen. Touchscreen, everything that make it fast for time, have an earbud in there, so they can hear everything that's going on… Voice commands, ‘Can you repeat the address that I'm going to?’ And it'll show everything. Callback number, ‘Can you call for me to make sure that the victim is there or I'm at the right address?’ Or, ‘I need an interpreter.’ Stuff like that. So everything is hands free…I don't have to press my mic. I'm just going to voice activate and start talking, and the computer's going to dial the numbers for you. You can have it all on the screen on your windshield. Everything.” (LE-U-012)

As described in the quote, LE responders were often interested in user interfaces that would allow them to more simply and intuitively receive and act upon information. Technology such as windshield HUDs could provide LE responders with accurate real-time information as well as allow them to more intuitively coordinate the resources they need for incident response.

4 Conclusion and Future Work

As development of the NPSBN and new communication technology continues, it will become increasingly important to create and refine user requirements for new communication technology designed for first responders. This study was an initial step in this process, examining first responders’ problems with communication technology and their requested functionalities. Although first responders saw great potential for communication technology to assist them with incident response, ultimately they felt that their technology was often unreliable and did not include functionalities well-suited to the environments they work within and the tasks they perform. Results from this study suggest that developers and designers may improve communication technology by addressing first responders’ current problems in three key areas.

First, first responders wanted their communication technology to be reliable. Addressing coverage and connectivity issues could help ensure first responders do not lose communication during incidents, making them safer and also allowing them to continue to send and receive information vital to incident response success. Second, developers and designers must carefully consider the usability needs of each first responder discipline. Lightweight communication technology that also integrates multiple devices or features into one device could improve the physical burden placed on first responders. Additionally, improving user interfaces to prioritize simplicity could optimize how quickly and effectively first responders react during incidents and send and receive mission critical information. Finally, first responders wanted improved interoperability with other disciplines, agencies, and jurisdictions. Many mentioned that communication was a critical component for incident response success, but many felt limited by their communication technology's ability to quickly connect them with other disciplines and information. Communication technology that can quickly and easily connect first responders may improve how effectively and efficiently they are able to prepare for and respond to emergencies. Taken together, new communication technology that takes into account first responders' needs for reliability, usability, and interoperability may result in more accurate, efficient and effective incident response. First responders can spend less time focused on their communication technology, and more time preparing for and responding to incidents.

Although many participants were most interested in improvements to current communication technology, they were open to new technology. As stated by an FF responder, "I think there's room for [new technology] as long as it's durable and it's user-friendly. That's huge." (FF-U-025) Many participants specifically discussed the utility of new devices that can capture and transmit real-time video or images. They also were interested in HUDs and user interfaces that can provide information accurately and quickly during incident response. This suggests opportunities exist for developers and designers to create new communication technology to help first responders. However, first responders in this study stressed that their most immediate need was for improvements to the communication technology they currently have. As stated by an FF responder, "Instead of introducing all this extra new stuff let's, one, make sure what we have actually works better. And then two, let's not rely on it so much." (FF-U-042) This quote underscores that first responders are less interested in new devices, and more interested instead in new solutions to their problems and challenges.

As the NPSBN continues to be built out, there is great opportunity for researchers and developers to design innovative new communication technology for and with first responders. To provide solutions to first responders' communication technology problems and needs, it will be critical for future work to use the findings from this study to inform technology development. As highlighted in this study, it will also be important to continue employing user-centered design principles in research and development efforts, seeking to first understand first responders' requirements and context of use before developing communication technology for them [9,10]. Communication technology should be designed *for* as well as *with* first responders [8]. By including first responders in the product development process, new communication technology may address first responders' problems and ultimately provide them with technology addressing their needs. This may help ensure first responders will adopt and use new technology, and also may allow them to perform their duties with more effectiveness, efficiency, and satisfaction.

Acknowledgments. NIST would like to thank the many first responders, public safety personnel, and public safety organizations who graciously gave their time and input for this project.

References

1. Balachandran, K., Budka, K.C., Chu, T.P., Doumi, T.L., Kang, J.H.: Mobile responder communication networks for public safety. IEEE Commun. Mag. **44**(1), 56–64 (2006). doi:10.1109/MCOM.2006.1580933
2. Baldini, G., Karanasios, S., Allen, D., Vergari, F.: Survey of Wireless Communication Technologies for Public Safety. IEEE Commun. Surv. Tut. **16**(2), 619–641 (2014). doi:10.1109/SURV.2013.082713.00034
3. Middle Class Tax Relief and Job Creation Act of 2012, Public Law 112–96, 126 Stat. 156. http://www.gpo.gov/fdsys/pkg/PLAW-112publ96/pdf/PLAW-112publ96.pdf
4. Grandi, J.G., Ogren, M., Kopper, R.: An Approach to Designing Next Generation User Interfaces for Public-Safety Organizations. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 944–945. IEEE, Osaka, Japan (2019). doi:10.1109/VR.2019.8797895
5. Doumi, T., Dolan, M.F., Tatesh, S., Casati, A., Tsirtsis, G., Anchan, K., Flore, D.: LTE for public safety networks. IEEE Commun. Mag. **51**(2), 106–112 (2013). doi:10.1109/MCOM.2013.6461193
6. Kumbhar, A., Güvenç, İ.: A comparative study of Land Mobile Radio and LTE-based public safety communications. In: SoutheastCon 2015, pp. 1–8. IEEE, Fort Lauderdale, FL (2015). doi:10.1109/SECON.2015.7132951
7. Park, E., Kim, J.H., Nam, H.S., Chang, H.-J.: Requirement Analysis and Implementation of Smart Emergency Medical Services. IEEE Access **6**, 42022–42029 (2018). doi:10.1109/ACCESS.2018.2861711
8. Choong, Y.-Y., Dawkins, S., Furman, S., Greene, K.K., Prettyman, S.S., Theofanos, M.F.: Voices of First Responders – Identifying Public Safety Communication Problems: Findings from User-Centered Interviews, Phase 1, Volume 1. NIST Interagency or Internal Report (NISTIR) 8216, National Institute of Standards and Technology (2018). doi:10.6028/nist.Ir.8216
9. Hackos, J.T., Redish, J.: Chapter 2: Thinking about Users. In: User and Task Analysis for Interface Design, pp. 23–50. Wiley, New York (1998)
10. International Organization for Standardization (ISO): Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. ISO 9241-210:2019, (2019)
11. National Institute of Standards and Technology (NIST): Research Portfolios. https://www.nist.gov/ctl/pscr/research-portfolios

12. National Institute of Standards and Technology (NIST): User Interface/User Experience. https://www.nist.gov/ctl/pscr/research-portfolios/user-interfaceuser-experience

13. Theofanos, M.F., Choong, Y.-Y., Dawkins, S., Greene, K.K., Stanton, B., Winpigler, R.: Usability Handbook for Public Safety Communications: Ensuring Successful Systems for First Responders. NIST Handbook (HB) 161, National Institute of Standards and Technology (2017). doi:10.6028/NIST.HB.161

14. Greene, K.K., Dawkins, S., Prettyman, S.S., Konkol, P., Theofanos, M.F., Mangold, K., Furman, S., Choong, Y.-Y., Steves, M.P.: Voices of First Responders—Nationwide Public Safety Communication Survey Methodology: Development, Dissemination, and Demographics, Phase 2, Volume 1. NIST Interagency or Internal Report (NISTIR) 8288, National Institute of Standards and Technology (2020). doi:https://doi.org/10.6028/NIST.IR.8288

15. Greene, K.K., Dawkins, S., Choong, Y.-Y., Theofanos, M.F., Prettyman, S.S., Furman, S., Steves, M.: Characterizing First Responders' Communication Technology Needs: Towards a Standardized Usability Evaluation Methodology. In: Mattson, P.J., Marshall, J.L. (eds.) ASTM Symposium on Homeland Security and Public Safety: Research, Applications, and Standards. STP 1614, pp. 23–48. ASTM International, West Conshohocken, PA (2019). doi:10.1520/stp161420180048

16. Dawkins, S., Choong, Y.-Y., Theofanos, M.F., Greene, K.K., Furman, S., Steves, M., Prettyman, S.S.: Voices of First Responders – Examining Public Safety Communication Problems and Requested Functionality: Findings from User-Centered Interviews, Phase 1, Volume 2.1. NIST Interagency or Internal Report (NISTIR) 8245, National Institute of Standards and Technology (2019). doi:10.6028/nist.Ir.8245

17. Dawkins, S., Greene, K.K., Prettyman, S.S.: Voices of First Responders—Nationwide Public Safety Communication Survey Findings: Mobile Devices, Applications, and Futuristic Technology, Phase 2, Volume 2. NIST Interagency or Internal Report (NISTIR) 8314, National Institute of Standards and Technology (2020). doi:10.6028/NIST.IR.8314

18. Federal Emergency Management Agency (FEMA): Regions. https://www.fema.gov/about/organization/regions

19. Evarts, B., Stein, G.P.: NFPA's US Fire Department Profile 2018. Report, National Fire Protection Association, (2020).

20. Crooke, C.: Women in Law Enforcement. Community Policing Dispatch **6**(7), (2013). https://cops.usdoj.gov/html/dispatch/07-2013/women_in_law_enforcement.asp

21. National Highway Traffic Safety Administration's Office of Emergency Medical Services National 911 Program: Next Generation 911. https://www.911.gov/issue_nextgeneration911.html

Workshop on ns-3 – WNS3 2021 – ISBN: 978-1-4503-9034-7
Virtual Event, USA – June 23-24, 2021

NetSimulyzer: a 3D Network Simulation Analyzer for ns-3

Evan Black
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
evan.black@nist.gov

Samantha Gamboa
Associate, National Institute of Standards and Technology
Prometheus Computing LLC
Sylva, North Carolina, USA
samantha.gamboa@nist.gov

Richard Rouil
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
richard.rouil@nist.gov

ABSTRACT

The increased complexity of network protocols and scenarios simulated using ns-3 is making the verification of simulation correctness and the analysis of simulation outputs a challenging task. In this paper, we present a new and flexible visualization tool for ns-3, called NetSimulyzer, that can alleviate the workload of debugging, understanding, and demonstrating a scenario. The tool was conceived to easily integrate to any ns-3 scenario and provides core functionalities that are technology agnostic. NetSimulyzer provides mechanisms to track a variety of simulation elements, from topology and node mobility, to statistics and other data generated by the simulated network protocols. The collected information can be visualized using a 3D scene augmented with data visualization elements such as charts and logs. In this paper, we provide an overview of the architecture and functionalities of the tool, and we also illustrate its usability and versatility by visualizing scenarios provided in the standard ns-3 distribution.

CCS CONCEPTS

• **Networks → Network simulations**; • **Human-centered computing → Visualization systems and tools**.

KEYWORDS

ns-3, network visualization, 3D visualization, Qt

ACM Reference Format:
Evan Black, Samantha Gamboa, and Richard Rouil. 2021. NetSimulyzer: a 3D Network Simulation Analyzer for ns-3. In *2021 Workshop on ns-3 (WNS3 2021), June 23–24, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3460797.3460806

Disclaimer: Any mention of commercial products in this paper is for information only; it does not imply recommendation or endorsement by NIST.

1 INTRODUCTION

The ns-3 Network Simulator is a discrete-event simulator widely used in research and academia [9]. It can simulate a variety of communication network architectures and protocols. As open-source software, users can also extend or implement new models that can use the underlying simulation framework. A typical workflow when using ns-3 includes the creation and parameterization of a scenario, the execution of the simulation(s), and the analysis of the simulation output data. Users can enable the ns-3 logging system or use the standard outputs to monitor the progress of the simulation. In addition, users make use of the tracing system, the modules' built-in metric systems, or the data collection framework [10], to gather simulation data of interest.

Interpreting network performance has become more complex due to several factors, including the increasing number of devices, the complexity of the protocols, the integration of different technologies, and the deep connection between upper and lower layer performance. Thus, relying upon raw data outputs to debug, understand, and demonstrate a scenario is cumbersome. Users often rely on custom post-processing tools that are scenario and technology specific and difficult to generalize or scale.

Separate visualization software may be used to ease simulation interpretation. Such software allows ns-3 users to reproduce the simulation and display performance metrics in a user-friendly manner. Visualizers also facilitate verifying correct scenario implementation such as node topology, simulation environment, and event timelines, and allow users to study protocol behavior through state machines and message exchanges.

There are two visualizers included with ns-3: NetAnim [4] and PyViz [11]. NetAnim is an offline visualizer, i.e., it uses an eXtensible Markup Language (XML) trace file generated at the end of a simulation to replay events in a separate application. PyViz is an online visualizer, i.e., it runs together with the simulation allowing live representation and debugging of a scenario. Both visualizers can show 2D representations of the scenario topology and animate node mobility and network connections between nodes. NetAnim can animate packet exchange between nodes over the displayed connections, while PyViz can display link statistics (e.g., data rate). NetAnim software integrates some summary tabs that can show metadata captured during the simulation, e.g., packet flows, node counters, and routing tables.

While NetAnim and PyViz are very useful for visualizations that only require the above features, extending the information to trace and display data or metrics, other than the ones embedded in their corresponding modules, is not trivial. In previous work, we built a demonstration of Mission-Critical Push-To-Talk (MCPTT) protocol capabilities over Long Term Evolution (LTE) Device-to-Device (D2D) communications using PyViz [6]. In addition to the limited documentation, which delayed our assessment of the tool, the tight coupling between the visualizer and the ns-3 modules (e.g., python bindings) required a considerable amount of effort and programming expertise to augment PyViz and obtain satisfactory results. Another limitation of NetAnim and PyViz is the lack of support

Black, Evan; Gamboa Quintiliani, Samantha; Rouil, Richard A. "NetSimulyzer: a 3D Network Simulation Analyzer for ns-3." Presented at Workshop on ns-3 (WNS3), Gaithersburg, MD, US. June 23, 2021 - June 25, 2021.

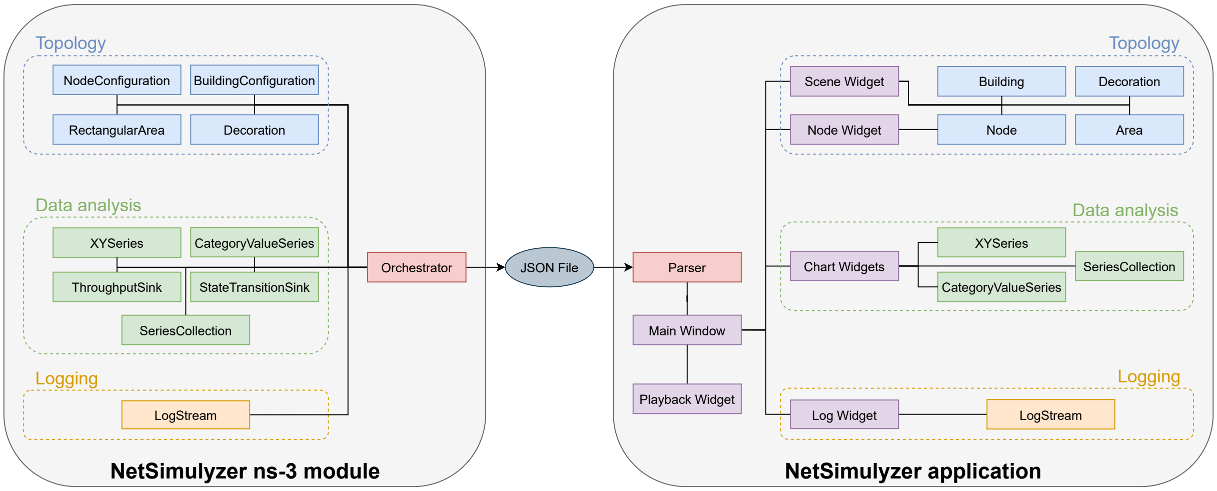


Figure 1: NetSimulyzer Architecture Overview

for 3D visualization even though ns-3 supports node positioning and mobility in 3D space. This likely prevents their effective use to support active research in aerial networks [2, 5, 12, 15] and millimeter wave communication [1, 18] where altitude and transmission angles play an important part.

Other users in the ns-3 community have implemented custom visualizers to suit their studies. For example, the authors in [17] describe a visualizer developed for the Institute of Electrical and Electronics Engineers (IEEE) 802.11ah ns-3 model. It is an online visualizer that can display a 2D representation of the topology, configuration, traffic statistics, and metrics of interest for 802.11ah (e.g., the slot usage) during simulation. In [1], the authors developed an offline visualizer to analyze beamforming training for the IEEE 802.11ad/ay ns-3 models using millimeter wave bands. This tool represents in 3D the scenario topology and environment, and the evolution of the beam steering configuration of the transmitters and receivers in the simulation. While these tools provide enhanced scenario visualization, they are currently limited to the specific technologies that the respective authors are studying. One open-source tool that aims to be generic is the United States Naval Research Labs (NRL)'s Scripted Display Tool (SDT) [16]. There are two versions of the tool: SDT, which provides 2D visualization like NetAnim and PyViz, and SDT3D, a Java-based application that provides 3D visualization by overlaying topology information on top of a map of the world. While it is possible to export an SDT3D compliant file from ns-3, it requires to use geographic coordinates to place the nodes on a map. Geographic coordinates are supported in ns-3 but not widely used. Furthermore, the terrain shown in the visualization is unlikely to represent what was modeled in the simulation unless the scenario is configured to use a terrain-aware propagation model such as Terrain Integrated Rough Earth Model (TIREM) available via contributed model to ns-3 [13]. Finally, SDT3D does not provide plotting capabilities to show network performance, thus requiring manual scripts to generate plots or use another library.

Due to the aforementioned limitations, we decided to implement a new visualizer that provides a 3D representation of the topology, core capabilities that are technology agnostic, and flexible simulation data collection and visualizations. Since the tool is meant to be used by other ns-3 users, it was also critical to make integration with existing and new scenarios as easy as possible.

The rest of the paper is organized as follows. Section 2 describes the structure and components of our tool and the functionalities they provide. In Section 3, we use several examples to demonstrate the capabilities and usability of the tool. Section 4 describes new functionalities that we are currently developing or plan to develop in the near future. Finally, Section 5 provides some concluding remarks.

2 NETSIMULYZER OVERVIEW

The NetSimulyzer is a tool designed to visualize and aid in understanding ns-3 scenarios of any size and any communication network technology used. NetSimulyzer is currently based on the offline approach and, as shown in Figure 1, it works by combining two separate entities: the NetSimulyzer ns-3 module, which reads the simulation data, and the NetSimulyzer application, which displays it. Both entities are open source and respectively available at [7] and [8].

The NetSimulyzer ns-3 module is used when configuring an ns-3 scenario and provides simple functions to specify the data to be collected during the simulation. The ns-3 module will generate a JavaScript Object Notation (JSON) file, which is used by the NetSimulyzer application to visually represent the simulation using different elements: a 3D scene that represents and animates simulation components in 3D, and multiple mechanisms to show the collected data in meaningful ways, e.g., with logs and interactive charts.

2.1 NetSimulyzer ns-3 Module

The NetSimulyzer ns-3 module is used to configure, collect, and export simulation information for display in the NetSimulyzer application. The module is designed to be modular, and the Orchestrator component is the only mandatory component that needs to be added to a scenario in order to use the tool. Every other component can be added based on what is relevant to the given scenario. Another design goal was the simplicity of integration; as such, the components typically require very few lines of code to add to existing ns-3 scenarios as shown in Section 3. In this section, we group the components currently present on the module by functionality and we describe their main characteristics and functions.

2.1.1 Orchestrator. The Orchestrator class is the base of the NetSimulyzer ns-3 module. It sets simulation-wide options, defines the output file path, and controls the output of tracked information to the JSON file. Every class of the module requires some reference to an Orchestrator, typically in the constructor. A single simulation may support several Orchestrator instances, allowing for several outputs from one simulation.

2.1.2 Topology. The module provides mechanisms to track the evolution of the simulation topology, supporting both items already found in ns-3, such as nodes and buildings, as well as additional items to further describe or enhance the environment and topology visualization, such as decorations and highlighted areas.

The class NodeConfiguration holds the visualization configuration of a node, such as its name and 3D model to use for display in the application, and monitors its position throughout the simulation. To begin tracking a node object for display, users aggregate a NodeConfiguration object onto it, and the module will automatically track the node for the rest of the simulation. The module will also automatically integrate with any of the ns-3 mobility models to easily track and output the node's location without additional code. The NodeConfiguration class supports many additional, optional properties to further refine the rendering of a node, such as configurable colors, height, orientation, offset, etc.

The NetSimulyzer ns-3 module is also capable of displaying buildings that are defined in the ns-3 scenario using a similar Application Programming Interface (API) to nodes, provided by the class BuildingConfiguration. To track a building object, users aggregate a BuildingConfiguration object onto it. Each building location, dimensions, and rooms are exported and will be displayed in the application as a solid or semi-transparent rectangular prism with planes dividing the rooms. The NodeConfigurationHelper and BuildingConfigurationHelper classes are also provided to allow simple configuration of many nodes and buildings at once.

To draw attention to locations of some significance in the simulation; the module provides the RectangularArea class. These areas are defined by an ns-3 rectangle and are displayed in the application with a border, fill, or both. The colors of the fill and border of an area can be configured in the scenario. Finally, to add purely visual enhancements to the topology scene, the module provides the Decoration class with a similar API to NodeConfiguration. This class alleviates the need to create nodes for props based on 3D models.

2.1.3 Data Analysis. Most of the network protocol information from an ns-3 simulation cannot be represented by the topology but instead must be captured via standard outputs or the ns-3 tracing system (i.e., by using trace sources and sinks).

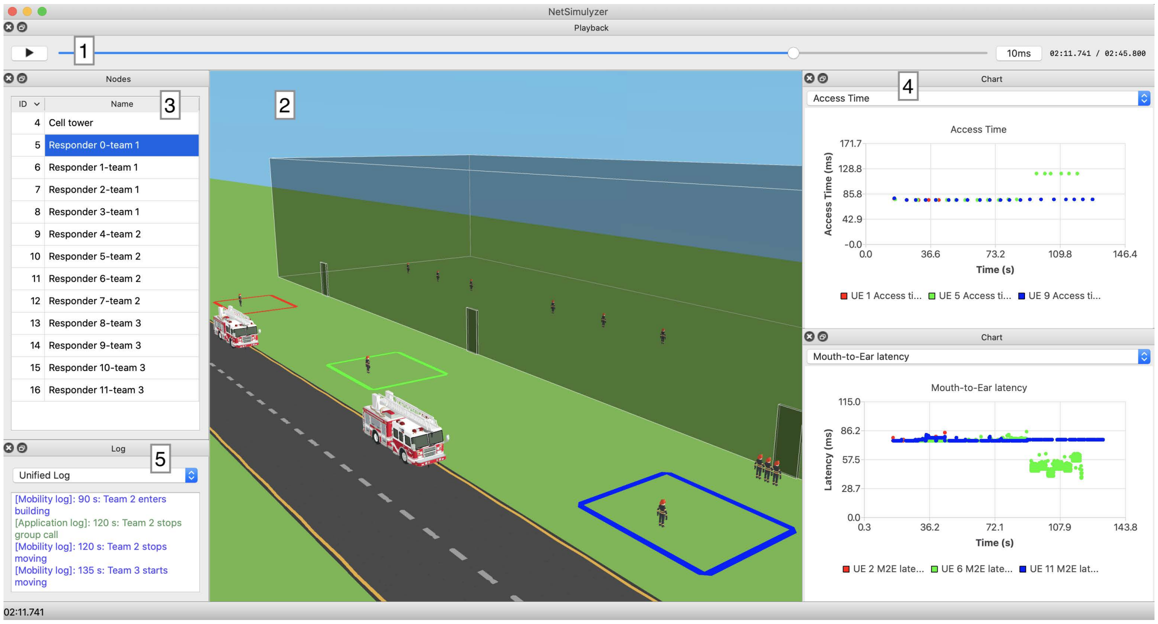
The NetSimulyzer ns-3 module currently provides three generic classes to collect such information from the simulations and to visualize it in the application using plots. The XYSeries class tracks information that can be expressed using two numeric coordinates, e.g., metrics over time, 2D positions, etc. By default, the module produces a plot per XYSeries object, and additionally, the SeriesCollection class can be used to group several XYSeries objects to be displayed in a single plot. The CategoryValueSeries class tracks numeric data organized into discrete, String categories, e.g., the state names of a state machine. Each class supports a number of properties that may be configured via the ns-3 attributes system or functions for configuring the plots. These properties include plot title, axis label, colors for the series, and the type of plot (lines, dots, connected dots). After an XYSeries or CategoryValueSeries is created, the data can be appended to it during simulation time. While there is no dependency on the ns-3 tracing system to generate statistics for the NetSimulyzer, this is typically used when instrumenting a scenario. For example, using a XYSeries object for collecting information, as shown later in the example of Section 3.1.

The module also provides two helper sinks to collect and process data to calculate throughput and display state machine changes in an easy way. The ThroughputSink tracks total data written by a model over a configurable period of time, calculates the throughput, and uses an XYSeries object to produce a plot that shows the throughput over time. The user can configure the interval at which the throughput is calculated as well as the unit for the Y-axis (e.g., bit/s, MB/s) based on expected throughput. This sink was designed to easily connect to packet transmission (Tx) and reception (Rx) trace sources defined in most applications. The StateTransitionSink was designed to track state changes and uses a CategoryValueSeries object to plot the changes against the time they occurred. This sink works with models that provide a traced value for their current state that uses states stored as Strings or as Enumerated types.

2.1.4 Logging. The LogStream class provides a mechanism, independent of ns-3's logging framework (NS_LOG), to output String messages during the simulation playback in the application. The LogStream class provides an API similar to std::cout and can be useful for displaying messages about the status of the simulation, such as marking the beginning of important events or indicating some failure condition has occurred. Each LogStream can be configured with a name and font color so that they can be clearly identified in the NetSimulyzer application during playback.

2.2 NetSimulyzer Application

The NetSimulyzer application is a lightweight, standalone, cross-platform, open-source, Qt application that replays an ns-3 simulation tracked by the NetSimulyzer ns-3 module. The application is a collection of optional widgets centered around the SceneWidget. The SceneWidget uses OpenGL to render hardware-accelerated 3D graphics for displaying the environment and network topology. The other optional widgets allow for control of the reproduction



ID	Name
4	Cell tower
5	Responder 0-team 1
6	Responder 1-team 1
7	Responder 2-team 1
8	Responder 3-team 1
9	Responder 4-team 2
10	Responder 5-team 2
11	Responder 6-team 2
12	Responder 7-team 2
13	Responder 8-team 3
14	Responder 9-team 3
15	Responder 10-team 3
16	Responder 11-team 3

```
[Mobility log]: 90 s: Team 2 enters building
[Application log]: 120 s: Team 2 stops group call
[Mobility log]: 120 s: Team 2 stops moving
[Mobility log]: 135 s: Team 3 starts moving
```

Figure 2: Example of the NetSimulyzer Application Interface

of the simulation and the display of the collected data using charts and logs.

In this section, we describe the main components of the NetSimulyzer application grouped by functionality so that the reader can see the correspondence with Figure 1 and Section 2.1. Also, we present in Figure 2 an example of the NetSimulyzer application reproducing a scenario. In the figure, we also show a breakdown of the interface components that will be referred throughout this section.

2.2.1 Parser. The parser reads the output file from the NetSimulyzer ns-3 module into models understood by the NetSimulyzer application. It also collects some metadata about the simulation, such as the furthest points in each direction and the last event time to determine the ground plane size and how long to make the playback timeline.

2.2.2 Main Window. The `MainWindow` holds all of the widgets and provides areas on the top, bottom, left, and right to move these widgets, referred to as “docks". Any widget, aside from the `SceneWidget`, may also be detached from the window and moved around freely.

2.2.3 Playback. Playback of the simulation can be controlled by the `PlaybackWidget` (item 1 in Figure 2). It allows for starting and stopping playback with either the play/pause button or pressing the `P` key. The user may also seek within the simulation by dragging the slider in the center of the widget forwards or backward. When the slider is moved, all the widgets are adjusted to the state that they were for the new time.

2.2.4 Topology. The traced topology from ns-3 is displayed in the widget at the center of the window, referred to as the `SceneWidget` (item 2 in Figure 2). It displays the nodes, buildings, decorations, and areas defined in the simulation and a ground plane, skybox, and minor lighting effects. The user may move or rotate the 3D scene using the mouse and keyboard. It is the only widget that may not be detached, moved, or disabled, although other widgets may be resized to cover it entirely.

The `NodeWidget` (item 3 in Figure 2) displays the list of nodes rendered by the application, with their node ID from ns-3, and the name set in the `NodeConfiguration`. The user may reorder the table by clicking on the column headings, and may change the view in the `SceneWidget` to center on a node by double-clicking on a record in the table.

2.2.5 Data Analysis. The application renders `XYSeries`, `CategoryValueSeries`, and `SeriesColection` objects from the ns-3 simulation into plots. These plots may be displayed on a `ChartWidget` (item 4 in Figure 2), added to the window by opening the “Window” menu and clicking “Add Chart.” Several `ChartWidget` instances may be created to compare data between plots easily. The user may adjust the view of the chart and focus on specific areas by clicking and dragging with the mouse or by adjusting the zoom level using the keyboard.

2.2.6 Logging. The `LogWidget` (item 5 in Figure 2) manages the `LogStreams` from the NetSimulyzer ns-3 module. Each individual `LogStream` may be selected by name from the drop-down at the top of the widget. There is also a “Unified Log”, which displays messages from every `LogStream` with the name from the ns-3 module appended in front of each message.

3 EXAMPLES

In this section, we use several examples provided in the standard ns-3 distribution to showcase the capabilities of the visualizer. These scenarios and their respective JSON output files are provided with the visualizer to assist new users as they familiarize themselves with integrating the NetSimulyzer module with ns-3 scenarios, and so that it is not necessary to run ns-3 simulations prior to operating the tool. In addition, a recorded presentation describing and walking through the scenario shown in Figure 2 is also available in [14].

3.1 Outdoor Random Walk

The first example demonstrates how easy it is to integrate the visualizer to an existing scenario to display the topology, add a plot, and use the log feature. To that extent, we selected the outdoor-random-walk-example scenario provided in the buildings module as it includes many buildings and one moving node. The code necessary to render Figure 3 is shown in Listing 1.

One powerful feature of the visualizer, which is to display the various buildings and moving nodes in the 3D scene widget, can be achieved by simply adding lines 15 to 23. We also show how to create and configure a log component and an XYSeries plot (lines 24 to 32). For the purpose of demonstration, both elements are used to export node position information, which occurs in the callback function defined at line 5. The user can then better understand the scenario layout by navigating through the topology and follow the node's movement after clicking the playback button. The resulting plot ends up representing a 2D trace of the node's trajectory, a capability also present in NetAnim.

3.2 WiFi Bianchi

We now present an example based on the wifi-bianchi scenario provided in the Wi-Fi module that was developed to validate the performance of the Wi-Fi ns-3 model with the theoretical model developed by Bianchi [3]. In this scenario, a number of devices in very close proximity communicate with each other over Wi-Fi. The scenario supports many parameters, including the Wi-Fi technology used (e.g., 802.11b, 802.11g), packet size, duration, and connection mode (i.e., infrastructure or ad-hoc). The scenario is also logging various information across different layers, such as when devices start to transmit or receive a packet at the physical layer, medium access control (MAC) layer, and application layer, or when devices need to execute the backoff procedures.

In order to study the effect of congestion and enhance the analysis, the distance between the center device and the other devices is set to 10 cm and the traffic start of each device is staggered throughout the first 75 % of the simulation time. A screenshot of the visualization is shown in Figure 4, which can be obtained with about 200 lines of code (the scenario itself being over 1200 lines of code). On the left side, the list of nodes is displayed, which for this example shows 10 stations. The log shows the time at which each station starts sending traffic. The charts selected show plots of the total MAC throughput received across all the devices, the traffic between station 7 (blue phone) and station 8 (green phone), and the backoff duration at station 7. The combo boxes associated with each figure could be used to look at traffic between any pair of nodes or view the congestion window information which is also available.

By looking at the various plots, the NetSimulyzer can be an effective tool to help students and researchers understand the impact of congestion in a Wi-Fi network. At time 8.89 s, station 7 starts to transmit. We can observe the increase in throughput in the top two figures, with station 7 sending at a rate of 12 Mbit/s (blue line in the second plot). Initially, all the traffic is received by node 8 (green line in the second plot). After the third station starts transmitting traffic, at time 34.52 s, the network saturates. Given that all the stations in the scenario have the same sensing configuration, they will each share the capacity equally. As such we see a gradual degradation of the traffic received by station 8, until about 3 Mbit/s, roughly 1/10 of the total capacity (as shown on the top figure). The bottom plot also shows the backoff time experienced by node 7. As more stations transmit, the medium gets busy for longer periods of time and stations have to backoff for longer periods of time.

3.3 LTE Radio Link Failure

The final example, shown in Figure 5, is based on the lena-radio-link-failure scenario available in the LTE ns-3 module. The scenario is designed to test the radio link failure and handover capabilities by having a User Equipment (UE) moving at a constant speed across the coverage areas of various eNodeBs. Inputs to the scenario include the number of eNodeBs, counters for the radio link failures, control of error models, and simulation time. The outputs are used mainly to capture the state of the Radio Resource Control (RRC) connection at the UEs and eNodeBs.

During the integration with the NetSimulyzer, three `LogStreams` were defined: one for the applications, one for events happening at the eNodeBs, and one for events at the UEs. On the figure, we can see that different colors are assigned to each log, making it easy to read. Plots were created for the RRC state, aggregated received throughput (i.e. combining both uplink and downlink), and the signal strength of the surrounding eNodeBs measured by the Reference Signal Received Power (RSRP). Figure 5 also demonstrates the NetSimulyzer's ability to rearrange the layout, with the RRC state plot placed on the left of the 3D layout since it requires more space.

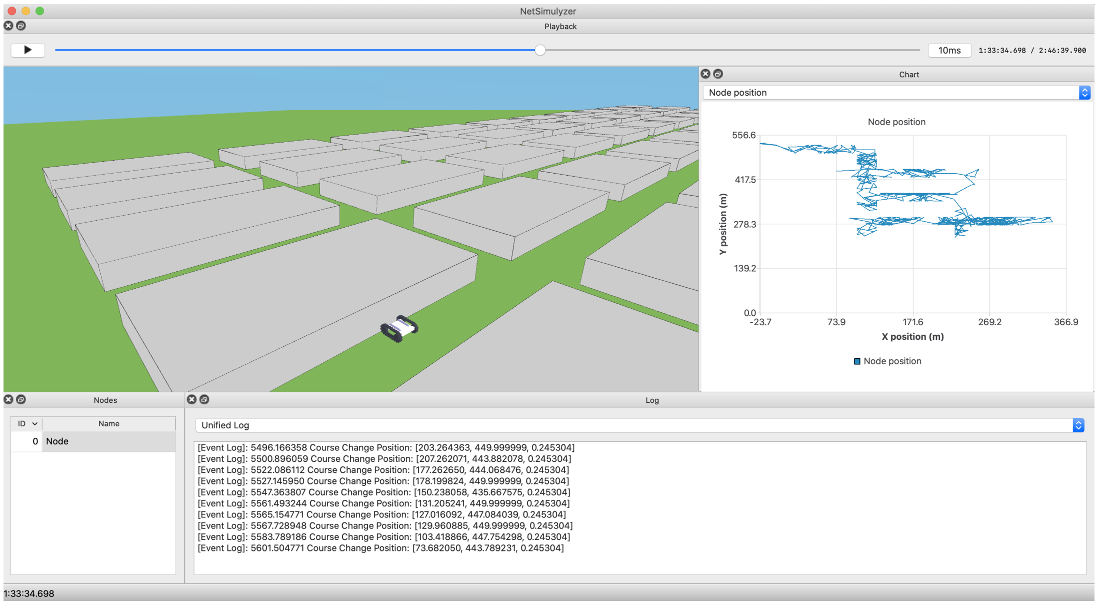
The selected plots show that the UE initially connects to cell 1 and the aggregated throughput is about 9 Mbit/s. As the UE moves away from cell 1, the signal strength becomes weaker, as shown by the green line in the RSRP plot. Consequently, the throughput decreases as the eNodeB performs link adaptation and reduces the Modulation and Coding Scheme (MCS) used for the UE. At 14.47 s, a radio link failure occurs, triggering a scan and a handover to cell 2. The RRC state diagram shows a transition to the CONNECTED_PHY_PROBLEM state before conducting a cell search and then re-establishing a connection. After the handover takes place, the UE is able to successfully send and receive packets, with increasing throughput as the UE gets closer to cell 2.

4 FUTURE WORK

As shown in the previous sections, the NetSimulyzer supports a large set of features to visualize and analyze ns-3 simulations. In this section, we introduce some new features currently in development or in our roadmap.

```
#include <ns3/netsimulyzer-module.h>
// ...
//Define callback function to track node mobility
void
CourseChanged (Ptr<netsimulyzer::XYSeries> posSeries, Ptr<netsimulyzer::LogStream> eventLog, std::string context, Ptr<const
      MobilityModel> model)
{
  const auto position = model->GetPosition ();
  //Write coordinates to log
  *eventLog << Simulator::Now ().GetSeconds () << " Course Change Position: ["
             << position.x << ", " << position.y << ", " << position.z << "]\n";
  //Add data point to XYSeries
  posSeries->Append (position.x, position.y);
}
// ...
auto orchestrator = CreateObject<netsimulyzer::Orchestrator> ("output.json");
//Use helper to define model for visualizing nodes and aggregate to Node object
netsimulyzer::NodeConfigurationHelper nodeHelper{orchestrator};
nodeHelper.Set ("Model", StringValue ("models/land_drone.obj"));
nodeHelper.Install(nodes);
//Use helper to configure buildings and export them
netsimulyzer::BuildingConfigurationHelper buildingHelper{orchestrator};
for (auto building = buildingVector.begin(); building != buildingVector.end(); building++)
  buildingHelper.Install(*building);
//Create a LogStream to output mobility events
Ptr<netsimulyzer::LogStream> eventLog = CreateObject<netsimulyzer::LogStream> (orchestrator);
eventLog->SetAttribute("Name", StringValue ("Event Log"));
//Create XYSeries that will be used to display mobility (like 2D plot)
Ptr<netsimulyzer::XYSeries> posSeries = CreateObject <netsimulyzer::XYSeries>(orchestrator);
posSeries->SetAttribute ("Name", StringValue("Node position" ));
posSeries->SetAttribute ("LabelMode", StringValue("Hidden"));
posSeries->SetAttribute ("Color", netsimulyzer::BLUE_VALUE);
posSeries->GetXAxis ()->SetAttribute ("Name", StringValue("X position (m)"));
posSeries->GetYAxis ()->SetAttribute ("Name", StringValue("Y position (m)"));
//Tie together the callback function, LogStream, and XYSeries
Config::Connect ("/NodeList/*/$ns3::MobilityModel/CourseChange", MakeBoundCallback (&CourseChanged, posSeries, eventLog));
```

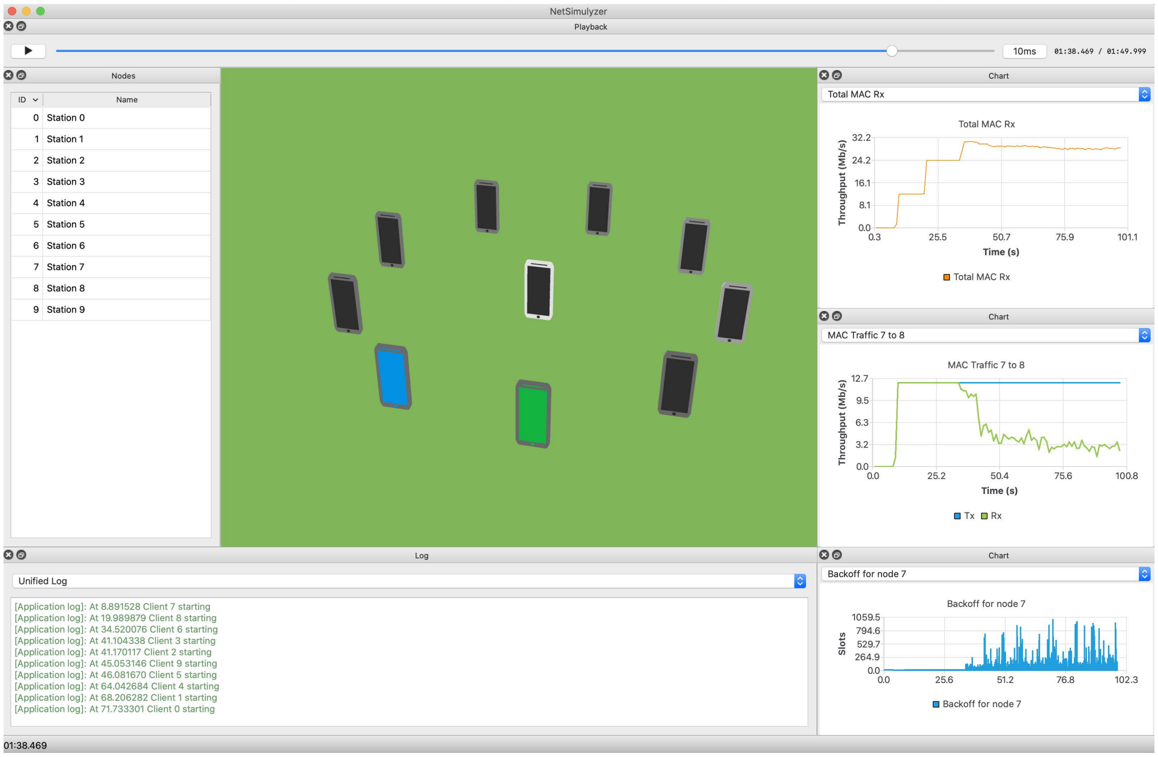
Listing 1: Code Needed to Visualize outdoor-random-walk-example as Shown in Figure 3



```
[Event Log]: 5496.166358 Course Change Position: [203.264363, 449.999999, 0.245304]
[Event Log]: 5500.896059 Course Change Position: [207.262071, 443.882078, 0.245304]
[Event Log]: 5522.086112 Course Change Position: [177.262650, 444.068476, 0.245304]
[Event Log]: 5527.145950 Course Change Position: [178.199824, 449.999999, 0.245304]
[Event Log]: 5547.363807 Course Change Position: [150.238058, 435.667575, 0.245304]
[Event Log]: 5561.493244 Course Change Position: [131.205241, 449.999999, 0.245304]
[Event Log]: 5565.154771 Course Change Position: [127.016092, 447.084039, 0.245304]
[Event Log]: 5567.728948 Course Change Position: [129.960885, 449.999999, 0.245304]
[Event Log]: 5583.789186 Course Change Position: [103.418866, 447.754298, 0.245304]
[Event Log]: 5601.504771 Course Change Position: [73.682050, 443.789231, 0.245304]
```

Figure 3: Visualization of outdoor-random-walk-example Scenario

Black, Evan; Gamboa Quintiliani, Samantha; Rouil, Richard A. "NetSimulyzer: a 3D Network Simulation Analyzer for ns-3." Presented at Workshop on ns-3 (WNS3), Gaithersburg, MD, US. June 23, 2021 - June 25, 2021.



ID	Name
0	Station 0
1	Station 1
2	Station 2
3	Station 3
4	Station 4
5	Station 5
6	Station 6
7	Station 7
8	Station 8
9	Station 9

```
[Application log]: At 8.891528 Client 7 starting
[Application log]: At 19.989879 Client 8 starting
[Application log]: At 34.520076 Client 6 starting
[Application log]: At 41.104338 Client 3 starting
[Application log]: At 41.170117 Client 2 starting
[Application log]: At 45.053146 Client 9 starting
[Application log]: At 46.081670 Client 5 starting
[Application log]: At 64.042684 Client 4 starting
[Application log]: At 68.206282 Client 1 starting
[Application log]: At 71.733301 Client 0 starting
```

Figure 4: Visualization of wifi-bianchi Scenario

We are currently developing additional widgets to display per-node network information, such as a list of `Applications` and `NetDevices`, and their associated information (e.g., device type, IP, and MAC addresses). The goal is to provide dynamic node status information relevant to the user.

We plan to expand the statistics capabilities. This will be done by providing additional trace sinks in the ns-3 module to include other frequently traced metrics, such as latency. Additional chart types, such as histograms and Cumulative Distribution Function (CDF) plots will also be developed.

We also plan to enhance the 3D scene by providing more network information such as antenna radiation patterns, the rendering of wired links, and traffic flow information. Other non-networking visual enhancements are also planned, such as support for text rendering to display node names, adding a 3D compass to help the user navigate through a scenario, or being able to select a node/link directly from the scene. This also includes expanding the catalog of available 3D object models to include more generic models of simple shapes and some network-specific models such as cell towers and Wi-Fi router models so users do not have to worry about creating or sharing their own models.

Finally, we also have plans to add an online mode, allowing visualization while the scenario is currently running, similar to the functionality provided by PyViz, but with the ease of integration provided by NetSimulyzer.

5 CONCLUSION

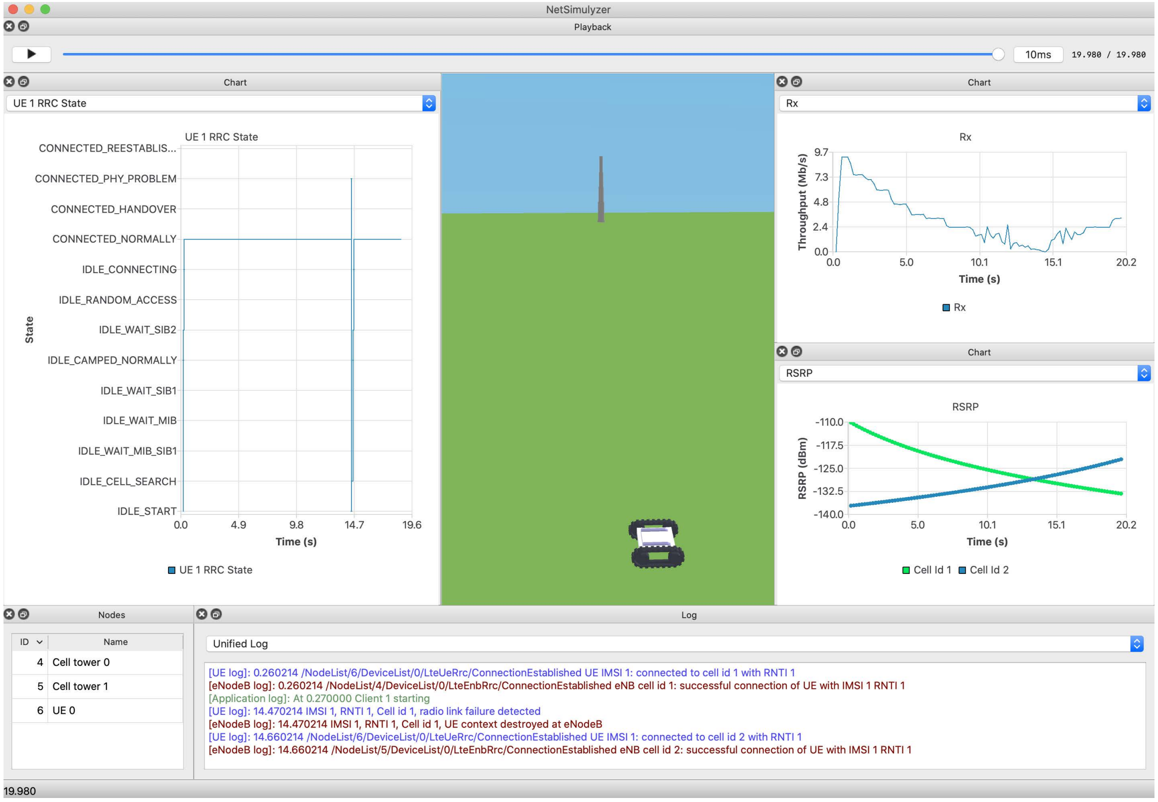
In this paper we presented a new open-source visualizer for ns-3 to fill the gaps of other existing visualization and animation tools. The 3D scene renderer provides the user with the ability to navigate through the topology and follow nodes' movements. The flexible plotting architecture allows users to select and configure the plots they want to show, and extend the capabilities to other types of plots. Using several examples, we demonstrated how this visualizer can help researchers understand their scenarios and share their results with the community. Finally, we discussed on-going work to further extend the supported features, which we hope can be done in collaboration with the ns-3 user community.

ACKNOWLEDGEMENTS

The authors would like to thank Tom Henderson for providing technical insights on ns-3 architecture and feedback throughout the development of this tool.

REFERENCES

[1] Hany Assasa, Joerg Widmer, Tanguy Ropitault, and Nada Golmie. 2019. 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* (Florence, Italy) *(WNS3 2019)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/3321349.3321354



ID	Name
4	Cell tower 0
5	Cell tower 1
6	UE 0

```
[UE log]: 0.260214 /NodeList/6/DeviceList/0/LteUeRrc/ConnectionEstablished UE IMSI 1: connected to cell id 1 with RNTI 1
[eNodeB log]: 0.260214 /NodeList/4/DeviceList/0/LteEnbRrc/ConnectionEstablished eNB cell id 1: successful connection of UE with IMSI 1 RNTI 1
[Application log]: At 0.270000 Client 1 starting
[UE log]: 14.470214 IMSI 1, RNTI 1, Cell id 1, radio link failure detected
[eNodeB log]: 14.470214 IMSI 1, RNTI 1, Cell id 1, UE context destroyed at eNodeB
[UE log]: 14.660214 /NodeList/6/DeviceList/0/LteUeRrc/ConnectionEstablished UE IMSI 1: connected to cell id 2 with RNTI 1
[eNodeB log]: 14.660214 /NodeList/5/DeviceList/0/LteEnbRrc/ConnectionEstablished eNB cell id 2: successful connection of UE with IMSI 1 RNTI 1
```

Figure 5: Visualization of lena-radio-link-failure Scenario

[2] Oscar G. Bautista and Kemal Akkaya. 2020. Extending IEEE 802.11s Mesh Routing for 3-D Mobile Drone Applications in ns-3. In *Proceedings of the 2020 Workshop on ns-3* (Gaithersburg, MD, USA) *(WNS3 2020)*. Association for Computing Machinery, New York, NY, USA, 25–32. https://doi.org/10.1145/3389400.3389406

[3] Giuseppe Bianchi. 2000. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications* 18, 3 (2000), 535–547. https://doi.org/10.1109/49.840210

[4] NetAnim. 2017. https://www.nsnam.org/wiki/NetAnim.

[5] Ben Newton, Jay Aikat, and Kevin Jeffay. 2015. Simulating Large-Scale Airborne Networks with ns-3. In *Proceedings of the 2015 Workshop on ns-3* (Barcelona, Spain) *(WNS3 2015)*. Association for Computing Machinery, New York, NY, USA, 32–39. https://doi.org/10.1145/2756509.2756514

[6] NIST. 2017. Clear Talk for First Responders. NIST Modeling Tool to Help Advance Cellular Emergency Communications. https://www.nist.gov/news-events/news/2017/10/clear-talk-first-responders.

[7] NIST. 2021. NetSimulyzer ns-3 Module, v1.0.1. https://github.com/usnistgov/NetSimulyzer-ns3-module.

[8] NIST. 2021. NetSimulyzer Visualizer Application, v1.0.1. https://github.com/usnistgov/NetSimulyzer.

[9] ns-3 Network Simulator. 2021. https://www.nsnam.org/.

[10] L. Felipe Perrone, Thomas R. Henderson, Mitchell J. Watrous, and Vinícius Daly Felizardo. 2013. The Design of an Output Data Collection Framework for ns-3. In *2013 Winter Simulations Conference (WSC)* (Washington, DC, USA). 2984–2995. https://doi.org/10.1109/WSC.2013.6721666

[11] PyViz. 2015. https://www.nsnam.org/wiki/PyViz.

[12] Paulo Alexandre Regis, Suman Bhunia, and Shamik Sengupta. 2016. Implementation of 3D Obstacle Compliant Mobility Models for UAV Networks in ns-3. In *Proceedings of the Workshop on ns-3* (Seattle, WA, USA) *(WNS3 2016)*. Association for Computing Machinery, New York, NY, USA, 124–131. https://doi.org/10.1145/2915371.2915384

[13] RemCom. 2021. EMPIRE Shim. https://apps.nsnam.org/app/empire/.

[14] Richard Rouil, Evan Black, Samantha Gamboa, Wesley Garey, and Thomas Henderson. 2020. Simulation and Visualization of Public Safety Incidents. https://www.nist.gov/ctl/pscr/simulation-and-visualization-public-safety-incidents.

[15] Benjamin Sliwa, Manuel Patchou, Karsten Heimann, and Christian Wietfeld. 2020. Simulating Hybrid Aerial- and Ground-Based Vehicular Networks with ns-3 and LIMoSim. In *Proceedings of the 2020 Workshop on ns-3* (Gaithersburg, MD, USA) *(WNS3 2020)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3389400.3389407

[16] US Naval Research Laboratory. 2021. Scripted Display Tool (SDT), a 3D Network Visualization Tool, v2.3. https://github.com/USNavalResearchLaboratory/sdt.

[17] Amina Šljivo, Dwight Kerkhove, Ingrid Moerman, Eli De Poorter, and Jeroen Hoebeke. 2018. Interactive Web Visualizer for IEEE 802.11ah ns-3 Module. In *Proceedings of the 10th Workshop on ns-3* (Surathkal, India) *(WNS3 2018)*. Association for Computing Machinery, New York, NY, USA, 23–29. https://doi.org/10.1145/3199902.3199904

[18] Tommaso Zugno, Michele Polese, Natale Patriciello, Biljana Bojović, Sandra Lagen, and Michele Zorzi. 2020. Implementation of a Spatial Channel Model for ns-3. In *Proceedings of the 2020 Workshop on ns-3* (Gaithersburg, MD, USA) *(WNS3 2020)*. Association for Computing Machinery, New York, NY, USA, 49–56. https://doi.org/10.1145/3389400.3389401

Black, Evan; Gamboa Quintiliani, Samantha; Rouil, Richard A. "NetSimulyzer: a 3D Network Simulation Analyzer for ns-3." Presented at Workshop on ns-3 (WNS3), Gaithersburg, MD, US. June 23, 2021 - June 25, 2021.

STEM-in-SEM: A Re-Emerging Material Measurement Approach*

Robert R. Keller
National Institute of Standards and Technology, Boulder, CO 80305

Analytical STEM-in-SEM has undergone a striking resurgence in terms of both methodology development and applications over the past 10 to 15 years, driven in part by the significant technological potential promised by low-dimensional structures such as nanoparticles and 2D materials. Key to maximizing the quantifiable information from a structure by electron microscopy is extraction of as much information as possible from every electron. To this end, characteristic mean free paths should be comparable to the size of the probed volume. A decrease in incident electron beam energy necessarily leads to shorter mean free paths for both elastic and inelastic scattering. This implies that for a given beam current within a small volume of material, use of a lower beam energy will produce more electron scattering, thereby increasing information content for potential analysis. For example, 20 keV electrons show elastic mean free paths in the range of a few tens of nanometers, while the corresponding path lengths for 200 keV electrons range from a few tens to a few hundreds of nanometers.

The foundational concepts for performing measurements by sending a relatively low energy ($\lessapprox$ 50 keV) electron beam through a material originated with the very early work of von Ardenne in 1938 [1], where a 23 keV beam was transmitted through a ZnO crystal, revealing microstructural detail with an approximate resolution of 40 nm in an early form of bright-field STEM imaging. The first STEM-in-SEM system that appeared commercially became available in 1966 [2]. However, despite the early successes and demonstrations of the utility of STEM-in-SEM imaging through the 1970s [3], relatively narrow views of the analytical utility of SEM were maintained for many years [4].

During the mid-2000s, examples of STEM-in-SEM for characterization of polymers [5] and high-resolution imaging of carbon nanotubes [6] demonstrated the analytical power of this measurement approach for samples considered to be difficult to characterize due to their low atomic numbers. The use of even relatively simple transmission detection schemes such as electron conversion detectors or segmented chip configurations resulted in striking, high-contrast imaging.

The current STEM-in-SEM resurgence has also been driven by recent advances in hardware capabilities that apply equally well to a SEM platform and a TEM platform. For example, the appearance of in-lens instruments [6] and transmission diffraction capabilities [7] has significantly strengthened the analytical power of the scanning electron microscope. In this context, NIST initiated several years ago a program to develop the measurement components that would be critical to the development of a fully analytical transmission scanning electron microscope, with the target of establishing metrology aimed at and optimized for quantifying individual nanostructures. The approach is to take concepts found in high-energy STEM and then adapt and integrate onto a commercial SEM platform.

Figure 1 shows a conceptual view of a future STEM-in-SEM system capable of simultaneous capture of numerous signals. A key feature of this vision is the generation of a multi-dimensional data stack comprising various forms of imaging, diffraction, and spectroscopy from each individual beam position in the raster scan, along with development of correlative component pairs. To date, NIST has made progress in the realms of electron diffraction [7], angularly selective imaging [8], and 4D STEM [9]. Continued progress toward such a STEM-in-SEM system will require advances in elemental and/or electronic spectroscopies and multivariate hyperspectral methods.

Keller, Robert. "STEM-in-SEM: A Re-Emerging Material Measurement Approach." Presented at Microscopy and Microanalysis (M&M 2021). August 01, 2021 - August 05, 2021.

References:
[1] von Ardenne, M., *Z. Phys*, 1938, (**108**), 553-572.
[2] Klein, T., Buhr, E., Frase, C.G., *Adv. Imag. Electr. Phys.*, 2012, (**171**), 297-356.
[3] Wells, O.C., *Scanning Electron Microscopy*, 1974, McGraw-Hill, New York, 8-10.
[4] Hawkes, P.W., *Ultramicr.*, 2010, (**110**), 1101-1113.
[5] Guise, O., Strom, C., Preschilla, N., *Microsc. Microanal.*, 2008, (**14(Suppl. 2**), 678-679.
[6] Van Ngo, V., Hernandez, M., Roth, B., Joy, D.C., *Microsc. Today*, 2007, (**15**), 12-17.
[7] Keller, R.R., Geiss, R.H., *J. Microsc.*, 2012, (**245**), 245-251.
[8] Holm, J., Keller, R.R., *Ultramicrosc.*, 2016, (**167**), 43-56.
[9] Caplins, B.W., Holm, J.D., White, R.M., Keller, R.R., *Ultramicrosc.*, 2020, (**219**), 113137.

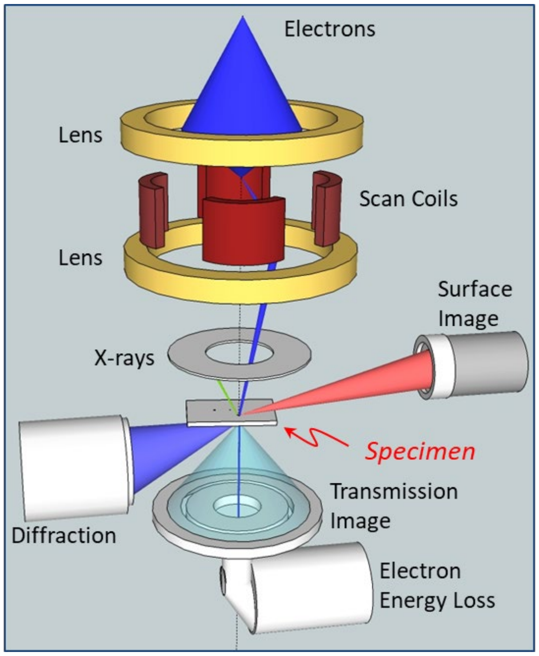


Figure 1 - Conceptual view of STEM-in-SEM system with full, simultaneous imaging, diffraction, and spectroscopy capabilities.

NIST SP 1283
August 2022

DETERMINING TEMPORAL EVOLUTION OF MASS LOSS PROFILES FROM MOCK-UPS OF STRUCTURAL COMPONENTS

Samuel L. Manzello[1] and Sayaka Suzuki[2]
[1]National Institute of Standards and Technology (NIST), USA
[2]National Research Institute of Fire and Disaster (NRIFD), Japan

1. INTRODUCTION

Wildland fires that spread into urban areas, termed Wildland-Urban Interface (WUI) fires, are becoming more and more common across multiple locations of the world. The 2018 WUI fires in the US state of California demonstrated the shear destruction that WUI fires are capable of by destroying more than 18,800 structures and resulting in multiple fatalities [1]. In 2020, the situation grew even worse, with multiple fires reported all over the Western US. WUI fires are one example of large outdoor fires. In Japan, urban fires have been the main large outdoor fire problem for centuries.

An important component in rapid spread of large outdoor fires is the production or generation of new, far smaller combustible fragments from the original fire source referred to as firebrands. Firebrands signifies any hot object in flight that are capable to ignite other fuel types. Firebrands are produced or generated from the combustion of vegetative and structural fuels. Firebrand processes include generation, transport, deposition, and ignition of various fuel types, leading to fire spread processes at distances far removed from the original fire source. While the term ember has been used to sometimes indicate the same connotation as firebrand, these terms are in fact slightly different. For WUI fires, the production of firebrands occurs from the combustion dynamics of vegetative and man-made fuel elements, such as homes. In the case of urban fires, the main culprit in firebrand production are homes.

The authors have shown that mock-ups of full-scale roofing assemblies may provide insights into firebrand generation from actual, full-scale roofing assemblies [2-3]. Here, the work has been extended to begin to understand more details of the combustion process of mock-ups of full-scale assemblies. The wind effect was explored on the temporal evolution of mass loss profiles for mock-ups of full-scale roofing assemblies.

2. EXPERIMENTAL DESCRIPTION

A series of experiments with mock-ups of full-scale roofing assemblies was performed in a wind facility in 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 % (two standard deviations) 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 oriented strand board and wood studs, with the dimensions of 0.61 m (W) x 0.61 m (H). As a first step, specific roofing treatments, such as roof tiles, were not applied. The mock-up assembly is half the height and width of the full-scale roofing assemblies (1.2 m by 1.2 m). The roofing angle was fixed at 25° (**see Figure 1**). The entire mock-up roofing assembly was placed on top of a load cell. Care was taken to protect the load cell from the heat generated from the combustion process. The assemblies were placed on top of gypsum board custom cut and sized to shield the load cells from the combustion process. The load cell was calibrated using a series of known mass samples prior to the experiments.

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 % (two standard deviations) 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.

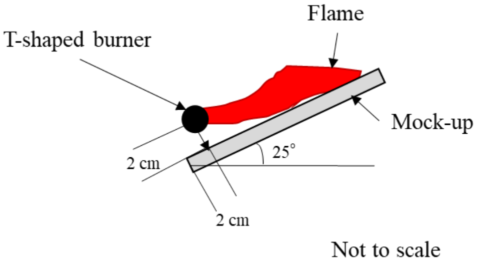


Figure 1 Schematic of the mock-up roofing assemblies used for the experiments. Side view is shown. Oriented strand board (OSB) was used as base sheathing.

Manzello, Samuel; Suzuki, Sayaka. "Determining Temporal Evolution of Mass Loss Profiles from Mock-Ups of Structural Components." Presented at 2021 Japan Association for Fire Science and Engineering Annual Symposium (JAFSE Annual Symposium 2021). May 29, 2021 - May 30, 2021.

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.

3. RESULTS AND DISCUSSIONS

Experiments were conducted for wind speeds of 6 m/s and 8 m/s. **Figure 2** displays the temporal evolution of mass loss for roofing assembly mock-ups at 6 m/s and 8 m/s. It is natural to assume that as the wind load was applied, this would result in artificial load on the mock-up assembly and this would influence the measured mass loss. To reduce this affect, the wind was turned on with the mock-up roofing assembly in place to determine the degree of actual offset. This offset was then subtracted from the measured mass loss profiles. The standard uncertainty in the temporal evolution of mass loss was approximately ± 5 % (two standard deviations).

The results are interesting as increases in attendant wind speed produce a faster reduction in the measured mass loss from the mock-up roofing assemblies. It is now possible to couple the temporal evolution of mass loss profiles to the measured size and mass distributions of firebrands liberated from the mock-up assemblies.

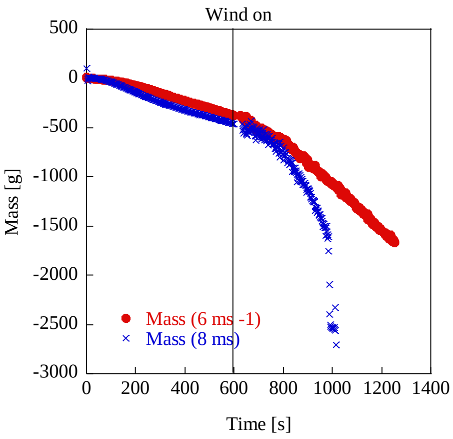


Figure 2 The temporal evolution of mass loss determined for mock-up roofing assemblies combusting under wind.

4. PRACTICAL IMPLICATIONS

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. Developing test standards for outdoor fire exposures presents significant challenges.

The main ignition mechanisms of structures from large outdoor 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. Radiant heat is a form of electromagnetic radiation that is emitted from sufficiently hot materials. Due to the combustion of fuels in large outdoor fires, any fuel type in close proximity to these combustion processes will experience radiant heat that may lead to ignition of these adjacent fuels.

Firebrand exposure is a dominant mechanism to structure ignition in large outdoor fires but global test methods for firebrands are lacking and are underway in ISO TC92/WG14 [4]. ISO TC92 has approved the development of an ISO standard firebrand generator for firebrand shower exposure. A current missing piece are simplified test methods that afford the ability to rank and rate building materials and vegetation for firebrand production.

5. SUMMARY

The wind effect was explored on the temporal evolution of mass loss profiles for mock-ups of full-scale roofing assemblies. The new experimental measurements have elucidated the influence of wind on the temporal evolution of mass loss for simplified structural components. Coupling these measurements with those of firebrand production from mock-ups of roofing assemblies are important to not only better understand the firebrand generation processes from structural fuels, but are also needed to develop new standard test methods to rank and rate building materials for the propensity for firebrand production.

5. REFERENCES

[1] Shulze, S., *et al.* (2020) Natural Hazards,104-901-905.

[2] Suzuki, S., Manzello, S.L. (2021) J. of Cleaner Production. 130: 135-140.

[3] Suzuki, S., and Manzello, S.L. (2020) Fuel 267:117154.

[4] ISO TC92/WG214 (Large Outdoor Fires and The Built Environment Working Group) https://www.iso.org/committee/50492.html (accessed on Feb 15, 2021).

Manzello, Samuel; Suzuki, Sayaka. "Determining Temporal Evolution of Mass Loss Profiles from Mock-Ups of Structural Components." Presented at 2021 Japan Association for Fire Science and Engineering Annual Symposium (JAFSE Annual Symposium 2021). May 29, 2021 - May 30, 2021.

MITIGATING IGNITION OF JAPANESE STYLE TILE ROOFING ASSEMBLIES

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

The 2016 Itoigawa City fire that occurred in Niigata, Japan, is an example of a recent large-scale urban fire in Japan. In the USA, large scale wildland-urban interface (WUI) fires are becoming more common, as evidenced by multiple WUI fires all over the Western USA in 2020.

Based on post-fire investigation surveys, it was reported that roofing assemblies were ignited by firebrand showers in both countries. Subsequent experimental work by the authors demonstrated these vulnerabilities. In the USA, it has also been demonstrated that tile roofing assemblies are known to be the ignition points by firebrands [1]. **Figure 1** demonstrates the dangers of firebrand showers to tile roofing assemblies.

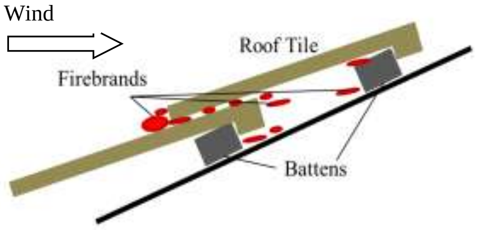


Figure 1 Firebrand penetration mechanisms under roof tiles; no specific tile type is shown [2]

A larger problem is what to do regarding the multitude of old roofing assemblies still present in Japan and the USA. It is not cost effective to take all these assemblies down and improve the ignition resistance of the underlayment. It is particularly important to the fire service, since in the event of a large outdoor fire event, ignitions are not often visible under roofing assemblies. The tiles obscure the smoldering ignitions that occur from the penetrated firebrands and often by the time the more hazardous smoldering to flaming ignitions occur, it may be too late to attempt to save the structure.

To this end, experimental work is underway to develop effective mitigation strategies that fire services may use in the event of large outdoor fire outbreaks. In this work, as a first step, wetting strategies were employed on roofing assemblies to determine if water application may reduce the penetration of firebrands under the tiles. The experiments made use of the authors firebrand generator technology coupled to well controlled wind facilities, as there is no point to conduct experiments under conditions that are not controlled.

2. EXPERIMENTS

Experiments were performed by using the full-scale Continuous-Feed Firebrand Generator. The experiments were conducted in the Building Research Institute's (BRI) Fire Research Wind Tunnel Facility (FRWTF). Experimental details are described [3] for Japanese style roofing assemblies. Only a short summary is provided here, as the procedures are similar, but the major changes were the application of water applied to the roofing assemblies just prior to start of the experiments (described in detail below).

The roofing assemblies were placed 2 m from the Continuous-Feed Firebrand Generator. A roof angle of 25 ° was used for all the roofing assemblies for direct comparison to prior work [2]. The overall dimensions of each roofing assembly were 1.2 m by 0.6 m. The wind speeds were 6 m/s and 9 m/s.

Prior to the experiments, water was applied to roofing assemblies. The premise of using water was to determine if wetting of roofing assemblies would result in less firebrand penetration under the tiles and thus mitigate ignition of the assemblies. The variables that were considered were wind speed, the total amount of water applied (in kg), and time to observe penetration. In these scoping experiments, the firebrand showers were intentionally selected to be in a glowing (smoldering) combustion state. The combustion state of the firebrands produced using the firebrand generator may easily be adjusted (smoldering, flaming, or combination of the two).

3. RESULTS & DISUSSIONS

In this work, two different wind speeds were considered, 6 m/s and 9 m/s. These wind speeds were selected to not only compare to prior work but are predicated on those measured in actual WUI fires or urban fires [3]. The amount of water applied before the firebrand generator and

Suzuki, Sayaka; Manzello, Samuel. "Mitigating Ignition of Japanese Style Tile Roofing Assemblies." Presented at 2022 Japan Association for Fire Science and Engineering Annual Symposium (JAFSE Annual Symposium 2021). May 29, 2021 - May 30, 2021.

wind were applied were 1 kg or 2 kg. **Figure 2** displays a typical experiment for an applied wind speed of 6 m/s and 2 kg of water applied to the tiles. It is important to note the water did not remain uniform on the surface due to the unevenness of the Japanese roof tiles, even though the water was evenly applied. During the the experiments, it was observed that water gradually dried out from wind and possibly accumulated firebrands absorbed water as they landed on the roof tile surface. Firebrands were also observed to be trapped in the water, on the surface, and extinguished.

After the desired firebrand exposure, either 10 min or 20 min, the roof tiles were removed and the number of firebrands that penetrated under the roof tiles was counted, for each case. Firebrands penetrated under the tiles, yet the battens were not charred, and no smoldering ignitions were observed as shown in **Fig. 3**.

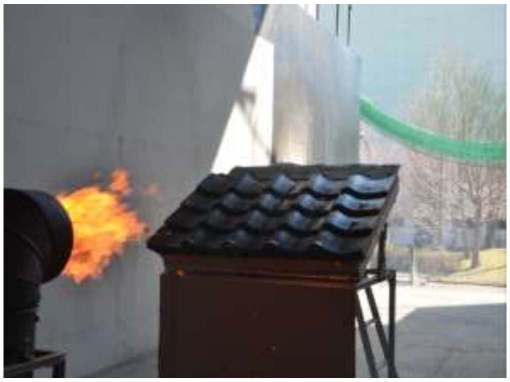


Figure 2 Image of a typical experiment. Applied wind speed of 6 m/s. 2 kg of water was sprayed onto the tiles just prior to the start of the experiments.

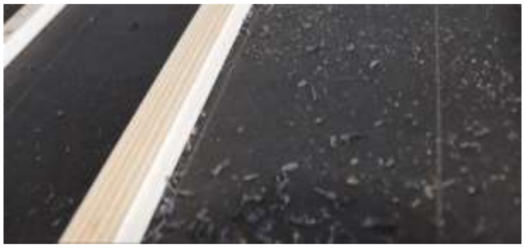


Figure 3 Battens and underlayment after the time have been removed. Applied wind speed of 6 m/s and 2 kg of water was applied before an experiment.

The number of firebrands that penetrated under the roof tiles were counted after each experiment. **Figure 4** shows the comparison of the number of firebrands penetrated under the tiles per unit area. With water applied before the experiment, the numbers decreased significantly, by 70 %. At the same time, **Fig.4** showed little difference to the reduction rate whether the amount of applied water was 1 kg or 2 kg. It is possible that a similar amount of water remained on the surface of roof tiles for both cases (1 kg or 2 kg of applied water).

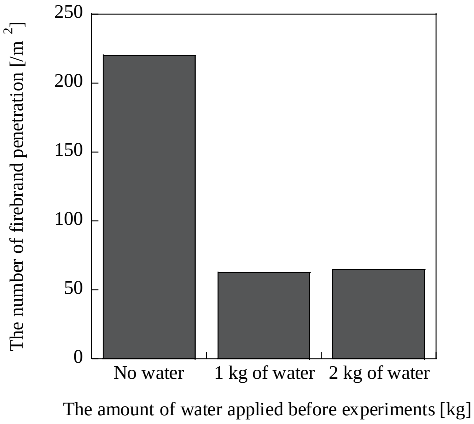


Figure 4 Comparison of the number of firebrands penetrated under roof under 6 m/s

4. PRACTICAL IMPLICATIONS

While exposure to wind-drive firebrand showers are a well-known vulnerability to ignition, to the authors knowledge, this is the first study to begin to investigate simple mitigation strategies for roofing assemblies**.** This study shows that the total amount of applied water may not be important, as no difference was observed from 1 kg to 2 kg. It would be important to know the minimum amount of water needed to prevent firebrands from igniting roofs for a certain time.

5. SUMMARY

In these experiments, scoping experiments were conducted to determine if wetting strategies were able to reduce the number of firebrands that penetrated the tile roofing assemblies and therefore mitigate ignition. Based on these early findings, none of the roofing assemblies ignited once water was applied prior to the experiments. The water acted to quench firebrands that landed on the roof tiles. More experimental work is required to verify these findings over a range of conditions.

5. REFERENCES

[1] Manzello, S.L., Fire Safety Science 11, (2014) 83-96.
[2] Suzuki, S., Manzello, S.L., Fire Saf. J., 91 (2017) 784-790.
[3] Suzuki, S. Manzello, S.L., Fire Technology, 56, (2020) 2315-2330

Suzuki, Sayaka; Manzello, Samuel. "Mitigating Ignition of Japanese Style Tile Roofing Assemblies." Presented at 2022 Japan Association for Fire Science and Engineering Annual Symposium (JAFSE Annual Symposium 2021). May 29, 2021 - May 30, 2021.

**Proceedings of the ASME 2021 International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
IDETC/CIE 2021
August 17-20, 2021, Virtual Conference, USA**

IDETC2021-71329

VISUALIZING MODEL-BASED PRODUCT DEFINITIONS IN AUGMENTED REALITY

Teodor Vernica[†♠], **Robert Lipman**[‡], **William Z. Bernstein**[‡*]
[†]Associate, Systems Integration Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA
[♠]Aarhus University, Department of Computer Science, 8200 Aarhus N, Denmark
[‡]Systems Integration Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA
Email: {teodor.vernica, robert.lipman, william.bernstein}@nist.gov

ABSTRACT

Augmented reality (AR) technologies present immense potential for the design and manufacturing communities. However, coordinating traditional engineering data representations into AR systems without loss of context and information remains a challenge. A major barrier is the lack of interoperability between manufacturing-specific data models and AR-capable data representations. In response, we present a pipeline for porting standards-based Product Manufacturing Information (PMI) with three-dimensional (3D) model data into an AR scene. We demonstrate our pipeline by interacting with annotated parts while continuously tracking their pose and orientation. Our work provides insight on how to address fundamental issues related to interoperability between domain-specific models and AR systems.

1 Introduction

Augmented Reality (AR) has become a valuable technology for manufacturing-based applications, including assistance in maintenance, process monitoring, and product assembly [1]. However, significant barriers exist to wider adoption of industrial AR including high development costs and fundamental lack of interoperability [2]. Aimed at achieving data interoperability for Smart Manufacturing Systems (SMS), the “digital thread” is conceptually useful for coordinating, aligning, and registering disparate data models across the product lifecycle including (but not limited to) product design, process planning, manufacturing execution, and part inspection [3].

Coordinating traditional engineering data representations into AR systems without loss of context and information remains a challenge. A major issue is the harmonization of standards within and across AR and SMS representations. For example, in previous work [4], we examined the integration issues between IndoorGML, a graph-based standard representation for modeling indoor spaces, with MTConnect, a standard for semantic interoperability of manufacturing assets. Though we were able to successfully generate a meaningful AR scene, we found semantic inconsistencies at the data-field level that can only be addressed by the standards development organizations (SDOs) themselves.

In this work, we investigate the feasibility of porting standards-compliant product definitions, including product manufacturing information (PMI) annotations, into AR environments. Figure 1 provides an example of a model-based product definition, displayed in isometric view of a computer-aided design (CAD) model with various annotations. The annotations adhere to the American Society of Mechanical Engineers (ASME) Y14.5 standard [5], an authoritative guideline for the design language of geometric dimensioning and tolerancing (GD&T). Many manufacturers treat such representations as living documents for reference throughout the product lifecycle. Hence, it is critical to reference the original definition for additional uses.

To translate PMI information into AR, we developed a pipeline that automatically leverages the standard product model as a reference. We validate our approach by using several parts in an assembly. To conclude, we enumerate challenges faced in the integration and use of the models in AR.

Our primary use case is overlaying standards-based in-

[*] Address all correspondence to this author.

Vernica, Teodor; Lipman, Robert R.; Bernstein, William Z. ”Visualizing Model-based Product Definitions in Augmented Reality.” Presented at 2021 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2021). August 17, 2021 - August 20, 2021.

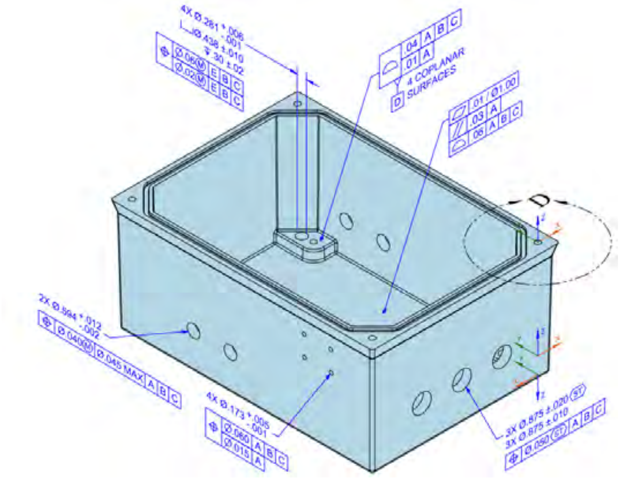


Figure 1: Example of a CAD model with PMI annotations[1].

spection data onto design information, including GD&T information, in an AR-compliant environment. Specifically, this work is a first step in coordinating our previous study [3] that mapped standards-compliant inspection data to design information through knowledge graphs. Integrating such perspectives into an AR-capable environment is essential for realizing highly scalable industrial AR.

2 Background and Related work

Industrial AR, or use of AR technologies in industrial practice, presents a number of domain-specific challenges. Many of those challenges relate to the compatibility of data representations across design/manufacturing-based use cases and AR presentation systems. For example, native design models are often represented as boundary-defined, three-dimensional (3D) models. To properly visualize such 3D models in AR, model simplification is required, often in the form of mesh-based representations. Translation leads to an inherent loss of information and fidelity, such as data associated with fully defined geometric features. One of the casualties of this process is GD&T information, a critical component for on-demand part inspection.

Below, we review the relevant data representations that facilitate translation of PMI into AR, related work for presenting part inspection data within virtual environments, and shortcomings of existing approaches.

2.1 Relevant data representations and standards

The STandard for the Exchange of Product model data (STEP) is a neutral representation of product data used for manufacturing [6]. STEP files facilitate interoperability between different CAD software and are used to represent PMI and other information vital for the smart manufacturing digital thread. STEP is maintained by the International Organization of Standardization[2] (ISO) and is actively developed to meet the requirements of the engineering community.

STEP AP242 [7], referred to as Managed Model Based 3D Engineering, covers the scopes of AP203 and AP214 and contains new capabilities for computer-interpretation of manufacturing and assembly information, including surface finish, manufacturing process information, and tolerances [8, 9].

The annotations in Fig. 1 are linked to a CAD model's features (e.g., edges, holes, and faces) to provide a formal definition of product geometry and specifications. PMI annotations include GD&T information and non-geometric data, e.g., surface texture specifications, surface finish requirements, process notes, material specifications, and welding symbols. Since GD&T is a symbolic language meant to communicate information about manufactured parts, standardization is vital for the presentation of annotations to be properly governed. ASME Y14.5-2009 [5] and ISO 1101:2012 [10] are the industry standards for the syntax and semantics of GD&T.

Though these standards exist, limited work addresses the porting of PMI into virtual environments, let alone AR, while still adhering to standard practices and guidelines.

2.2 Relating CAD software to AR engines

While AR is proving to have numerous industrial use cases, developing applications requires significant time investment. One fundamental barrier to quickening development time is the lack of interoperability between existing engineering data (e.g., CAD models) and the software used to develop AR applications.

AR applications are primarily developed in game engines such as Unity[3]. Game engines are traditionally used for video game development, but their powerful toolsets lend themselves to general software development including AR applications. However, game engines provide limited compatibility with engineering data, given that engineering work is curated in specialized software, with little overlap. By default, Unity does not support any of the widely used boundary representation (B-rep) CAD formats, such as STEP. This incompatibility can hinder the use of existing models in AR environments and therefore hinder experimentation and use of these new technologies.

With the increase in adoption of Industrial AR solutions [11], commercial efforts have emerged to bridge the gap between engineering data and game engines. These solutions can vary from model translation tools and importers, to stand-alone AR visualization platforms. Of important note are PiXYZ's studio[4]

[1] https://go.usa.gov/mGVm

[2] ISO TC184/SC4 is the subcommittee responsible for STEP.

[3] https://unity.com/

[4] https://www.pixyz-software.com/studio/

and plugin[5]. PiXYZ studio is a data preparation tool that enables the conversion and optimization of 3D CAD data to more lightweight tessellated representations, better suited for visualization purposes. PiXYZ plugin is a Unity plugin that provides a more direct integration and support for additional features, such as the ability to import point cloud data. Given the lack of native support, Unity has partnered with PiXYZ and endorses the plugin as the way to incorporate CAD data inside the game engine [12].

While commercial products, such as PiXYZ and others, provide robust and easy-to-use solutions for development, and have previously been successfully leveraged by researchers [13, 14], open solutions can provide additional benefits. Standards-based open-source solutions are desirable in many settings because of the transparency with which the data is handled. To the best of our knowledge, there are currently no established open workflows or pipelines that can achieve automated integration between 3D CAD data and Unity.

2.3 Use of PMI within AR

While limited, previous work has used AR as a tool for visualizing manufacturing data. Urbas et al. [13] propose a method for part inspection that aids users in the measurement process by contextually visualizing PMI information in AR. They make use of the PiXYZ plugin to import the CAD data into Unity, including graphical PMI annotations. However, the implementation of the AR application itself is not realized. In contrast, we showcase a means to achieve the same goal by leveraging standards-based open-source tools. Section 4.1 showcases our AR application, which validates our approach.

Fiorentino et al. [15] present the tangible digital master (TaDiMa) system. TaDiMa leverages markers embedded in technical drawings as tangible user interfaces to display Product Lifecycle Management (PLM) data, queried from a PLM database. Additionally, Fiorentino et al. showcase two methods for reducing annotation clutter and readability. Finally, they propose a number of potential use case scenarios for their system. They explore additional use cases of a similar AR system in another paper [16]. In our work, the focus is on the data integration and the automated extraction and visualization of the data defined in the models. To deal with annotation clutter, we employ a two-dimensional (2D) user interface consisting of a list of toggles for each view defined in the STEP file.

2.4 Takeaways from existing work

Realizing scalable and maintainable industrial AR experiences requires significant research and development. Such opportunities lie not only in fundamental research opportunities, such as understanding and improving asset tracking capabilities and enhancing worker engagement through more comfortable visualization modalities. However, to facilitate decision-making on the floor, it is still necessary to merge state-of-the-art AR technologies into existing engineering workflows.

To achieve this vision, a standards-based approach is necessary. In our approach, we rely on standards to overcome data interoperability challenges. Standards, specifically standard data representations, provide mechanisms for data interchange between disparate computer-aided engineering (CAE) software. Next, we describe our technical approach, focusing on specific design decisions for producing a proof-of-concept.

3 Technical approach

For design information, we assume that the STEP data model stores all as-designed geometry and other critical specifications, including GD&T annotations. Currently, AR engines and presentation systems cannot read STEP natively. Hence, it is necessary to translate the model into an AR-ready representation without losing context.

Model tessellation describes the process of translating a B-rep model into a triangular mesh, while model decimation refers to the process of reducing the number of polygons from an existing mesh [17]. Model tessellation and decimation is common for visualizing 3D models on lightweight devices, such as head-mounted displays. The magnitude of model decimation depends on a number of factors, including the computational power of the presentation device, e.g., tablet or head-mounted display.

For model tessellation, we leverage the NIST STP2X3D Translator[6], which inputs a STEP Part 21 (P21) file [18] (or a STEP instance file) and outputs an X3D file including annotations and views per the user's request. The exact part geometry is converted to faceted geometry for X3D, an ISO standard [19]. The NIST STEP File Analyzer and Viewer[7] (SFA) uses the translator for enhanced visualization of geometry and graphical PMI on a web browser. For the purpose of the presented pipeline, shapes in X3D are represented by coordinates that are connected to create lines and faceted surfaces using `<IndexedLineSet>` and `<IndexedFaceSet>`, respectively. We provide additional explanation in Sec. 4. The X3D format offers significant opportunities in its ability to support lightweight visualization and readiness to link domain-specific information to geometric features [20].

Our approach is fully standards-based and open-source. As a reference throughout the pipeline, we leverage STEP to represent part geometry and graphical PMI annotations. Exact part geometry is represented by free form surfaces and geometric primitives such as planes and cylinders. Examples of graphical PMI annotations for dimensions and geometric tolerances are shown in

[5] https://www.pixyz-software.com/plugin/

[6] https://github.com/usnistgov/STP2X3D/

[7] https://www.nist.gov/services-resources/software/step-file-analyzer-and-viewer

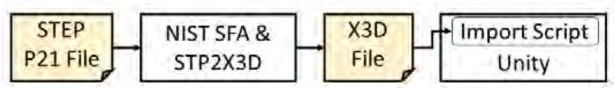


Figure 2: Implemented pipeline.

Fig. 1. Each annotation has a leader line that attaches it to the associated surfaces of the part. In the STEP file, graphical PMI is represented by lines and faceted surfaces.

4 Implementation details

After a STEP file is converted to a mesh representation using the STP2X3D Translator, the X3D file must be imported into Unity. While Unity lacks native support for the X3D format, files can be easily parsed using generic eXtensible Markup Language (XML) parsers, such as the one built in Microsoft's .NET platform on which Unity runs. Therefore, an additional import script is required. Figure 2 represents our implemented pipeline.

The import script is (1) attached to a game object in the scene hierarchy, (2) takes the X3D file generated by SFA as input and (3) generates the respective geometries using Unity functions and components according to Fig. 3. For example, the vertices encoded within the `<IndexedLineSet>` X3D element are drawn using Unity's `Line Renderer` component. To facilitate the manipulation of each drawn line, such as its scaling, translation and rotation, we convert each line into a mesh. `<IndexedFaceSet>` elements can directly generate a Unity mesh by using the encoded vertices together with the indexes encoded with the `coordIndex` attribute as the mesh triangles.

```
<IndexedFaceSet solid='false'
    coordIndex='16 17 18 -1 ... 16 18 19 -1'>
    <Coordinate DEF='coord100'
    point='4. 4.1 41.925 ... 6.476 4.1 42.30'/>
</IndexedFaceSet>
```

Listing 1: Example IndexFaceSet element.

Listing 1 shows an example `<IndexedFaceSet>` element. The element contains a `<Coordinate>` element as a child, which encodes coordinates through the `point` attribute. The coordinates can be used to generate a mesh in Unity with additional processing, as the mesh requires vertices as an array of type Vector3. The resulting array consisting of multiple Vector3 instances can then be passed to a new Unity mesh instance through its `vertices` property. To draw faces between vertices, indexes encoded in the `coordIndex` attribute of the `<IndexedFaceSet>` element are required. Indexes correspond to each Vector3 value, are separated by the -1 value, and can be passed to a Unity mesh through its *triangles* property, as an array of integers (excluding the -1 values).

Since the number of indexes of `<IndexedLineSet>` elements is not necessarily a multiple of 3 (e.g., `<IndexedLineSet coordIndex='0 1 -1'>`), it can not directly generate a mesh as the indexes do not always define triangles expected by the mesh. Instead, the indexes and point coordinates can be passed to a Line Renderer Unity component through the `SetPosition` method. The resulting line can be converted into a mesh using the `BakeMesh` method.

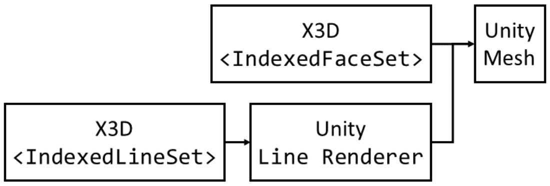


Figure 3: Mapping of elements from X3D to Unity.

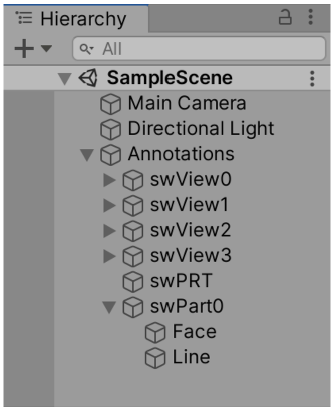


Figure 4: Example of generated Unity hierarchy.

Figure 4 shows an example scene hierarchy generated by the import script in Unity. In this example, the import script is attached to an empty game object named `Annotations`. Graphical PMI annotations are grouped in the X3D file by views, with each view containing different aspects of the part definition, such as tolerances and datum definitions. To preserve this structure, each view encoded in the X3D file generates a view game object in the scene hierarchy (`swView0`, `swView1`, etc.), as a child of the original object. The part geometry (`swPart0`) is encoded separately from the views and is also generated as a child of the original object, e.g., the `Annotations` game object in this case. Meshes generated from `<IndexedFaceSet>` and `<IndexedLineSet>` elements are then generated as children of each corresponding view or part geometry in the hierarchy (`Face`, `Line`). Preserving this structure allows views, part geometry and meshes to be manipulated individually within the game engine. Figure 5 illustrates the drawing of the part geometry with and without annotations. In this example, only the wire-

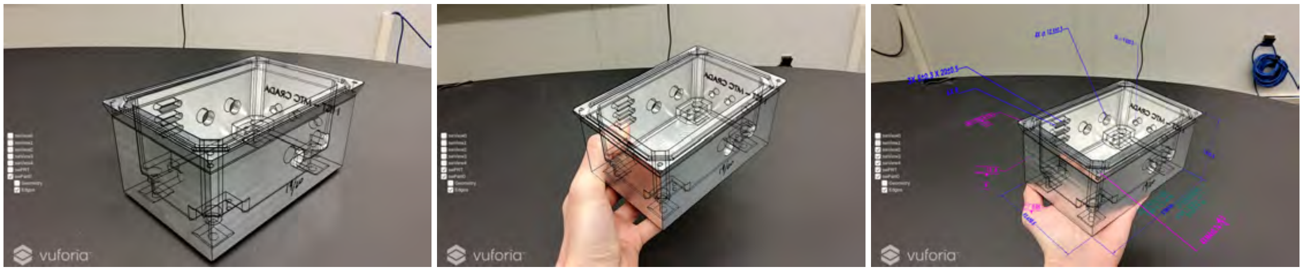


Figure 5: AR view of the wire-frame of the part geometry overlaid on the physical part.

frame (`Line`) of the geometry (`swPart0`) is superimposed over the physical piece, providing additional context.

The import script, is not a fully fledged X3D importer, as only relevant elements are being translated to equivalent Unity counterparts. But additional support can be easily added, e.g., `<Viewpoint>` X3D elements can be mapped to a Unity camera object, and `<DirectionalLight>` elements can generate a Unity object characterizing directional light.

4.1 Augmented Reality application

By leveraging the previously described pipeline (Fig. 2), we developed an AR application[8] which uses model-based tracking to register PMI annotations extracted from the as-designed CAD model, to their physical counterparts, as shown in Figs. 6 & 7.

The annotations are separated into views. Each view is represented by a different color, as encoded in the X3D file by SFA. Each view and the part geometry can be toggled on and off, using the user interface shown in Fig. 8. The toggles are automatically generated based on the currently tracked model. Figure 8a showcases two views simultaneously overlaid onto the box assembly. View 2 (magenta) represents datum and top hole definitions and View 3 (blue) represents the boundary and side hole definition. Figure 8b shows two additional views toggled on the same part: View 0 (green) displays notes and titles, and View 4 (cyan) shows the bottom hole definition.

To visualize the PMI annotations in AR, we leveraged PTC's Vuforia[10] framework due to its robust tool set. The recent addition of model-based tracking allows physical 3D objects to be used as tangible user interfaces [21]. This means that users can more naturally interact with digital information, such as CAD models, through a physical 3D representation of the model itself rather than through a 2D screen, with a mouse and keyboard as an input modality. Note that the AR framework itself is not tied to the importer, and thus other tracking solutions could be used.

[8] See demo at https://pages.nist.gov/CAD-PMI-Testing/NIST-AR-video.html

[9] https://pages.nist.gov/CAD-PMI-Testing/NIST-AR-plate.html

[10] https://www.ptc.com/en/products/vuforia

[11] https://pages.nist.gov/CAD-PMI-Testing/NIST-AR-cover.html

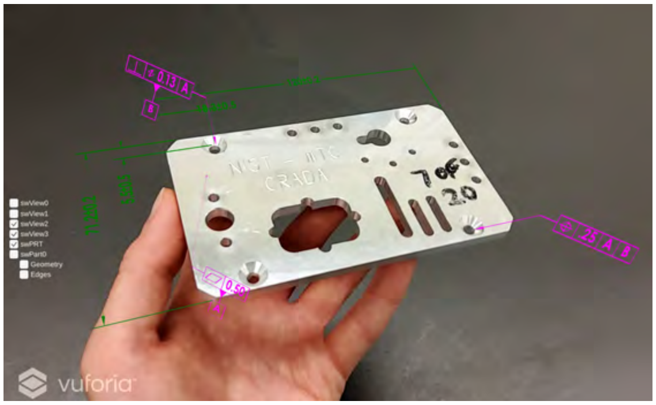


Figure 6: PMI annotations superimposed on the machined Plate[9] part from the NIST dataset. View 1 (magenta) shows datum definitions. View 2 (green) shows hole definitions.

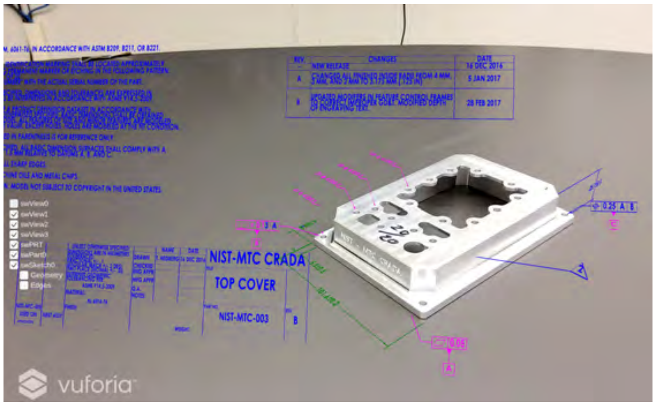
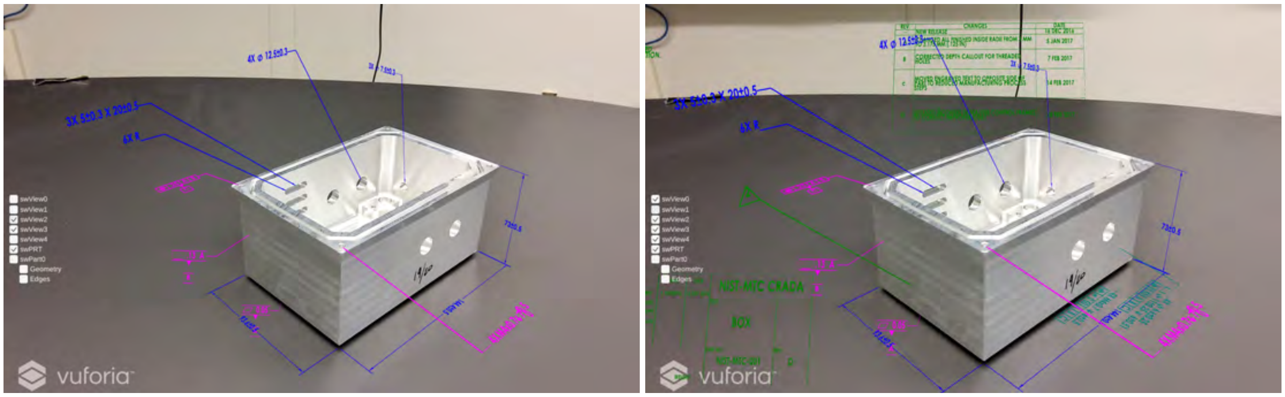


Figure 7: AR view of the Cover[11] part. View 1 (blue) presents notes and titles. View 2 (magenta) shows datum and hole definitions and View 3 (green) displays boundary definitions.



(a)

(b)

Figure 8: AR views of the Box[12] from the NIST Box Assembly dataset, with different views toggled.

To validate our approach, we imported and tracked three STEP files that make-up the NIST *Design, Manufacturing, and Inspection Data for a Box Assembly* dataset[13]. The publicly available dataset consists of three CAD Models and associated data collected during the fabrication process of the parts. The application was deployed to an Android tablet, running on the Android 10 operating system.

5 Pipeline issues

The presented workflow requires manual intervention when translating files from STEP to X3D using SFA. The process can be automated by using the command line version of SFA.

Additional human intervention, such as scaling and rotating, might be required once the model is imported inside the game engine. This is due to different units and coordinate system conventions used by software vendors. For example, Unity uses a left-handed y-up coordinate system, while another popular game engine, the Unreal Engine, uses a left-handed z-up convention. Additionally, by default, many 3D modeling programs use right-handed coordinate conventions. This means that a 3D model, created in one program, might be oriented and scaled differently depending on conventions used by the software to which it is imported. While these issues can be solved, it is usually done on a case-by-case basis.

Augmented reality tracking libraries generally provide a visual representation of the tracked model within the game engine. In doing so, digital augmentations can be positioned relative to a visual aid in the development process. Hence, it is possible that the visual aid would be sensitive to the same challenges mentioned previously. In other words, the visual aid could be oriented or scaled differently than the X3D encoding. Additional work might be required to correctly align the imported X3D geometry and annotations to the provided visual aid within the game engine. To minimize this overhead, our importer attempts to automatically scale the X3D part geometry and annotations if such a visual aid is present. Our algorithm compares the dimensions of the two meshes and computes the scale difference between them. The X3D geometry is then scaled up or down by a computed factor. Note that we do not modify the scale of the visual aid to avoid any impacts on the tracking process.

While this would help eliminate some potential overhead, additional manual intervention might still be required once the data is brought into Unity during the application creation process. This is out of scope for this work, and might be dependent on other factors, such as the tracking libraries used e.g., Vuforia's model target preparation process.

6 Future directions

Our proof-of-concept uncovers a number of research opportunities for a standards-compliant pipeline to be AR-ready with limited loss of information and context. Here, we present a number of research directions that we intend to further investigate.

- *Spatially contextualizing annotations with respect to the decimated part model*: In the native design model, PMI annotations are fixed to the features to which they characterize. However, once the model is tessellated into a mesh for AR presentation, annotations are no longer coupled to a closed-body feature. Instead, the placement of annotations relies on the coordinate-based placement of annotations. To ensure proper placement, it is necessary to properly segment the decimated model to contextualize annotation placement.
- *Coordinating contextual views of annotations*: The STEP

[12] https://pages.nist.gov/CAD-PMI-Testing/NIST-AR-box.html
[13] https://github.com/usnistgov/smstestbed/tree/master/tdp/mtc

data structure affords context views of annotations, which could be thought of layered data presentations based on user needs. The same capability could be leveraged to better handle AR-based presentation. The use of contextual views for AR is not well understood. As a result, there might exist opportunities for appending STEP with additional entities to facilitate better AR presentation.

- *Registering other digital thread data representation in the same spatial context*: Relating inspection data back to design models in AR is a desired use case. For example, non-destructive inspection (NDI) is leveraged for expensive parts, especially in high-mix, low-volume situations. Spatial registration between NDI data structures, such as those from X-ray computed tomography (XCT) scans for additive-manufactured parts, with native design models remains a challenge. Addressing automated registration of such data structures will facilitate unique AR use cases. Similar insights can be drawn from other manufacturing-related data, including point clouds derived from traditional probe-based measurement instruments and controller-reported data from machine tools.

For tracking parts in the presented AR application, we leveraged a commercial tracking toolkit. To accomplish a fully open-source solution, a recently released open-source tracking toolkit[14] could fulfill this requirement.

7 Conclusion

We presented an automated approach for linking detailed design data, including PMI, within an AR experience. The lack of automated methods for coordinating PMI into game engines is a primary barrier for its use in AR. This technical gap motivates our work. Our technique is rooted in standards and leverages open tools (when available) throughout the pipeline. Standards are critical for addressing interoperability challenges when coordinating domain-specific models with AR systems.

We envision this work will better facilitate integration of manufacturing information contextualized to the native design model. Proper registration of such data will enable more scalable industrial AR applications. As a result, future work will center around spatially contextualizing downstream data onto design models.

In summary, our pipeline simplifies AR scene creation for product model definitions. In the near term, the automated translation of native design models into Unity meshes is immediately useful for creating AR application prototypes more quickly. In the past, manual effort required to construct such functional prototypes has inhibited wider adoption of industrial AR. Our contributions can be leveraged across multiple use cases, including on-part inspection and assembly guidance.

Acknowledgements

We thank the NIST-internal and conference reviewers for their valuable feedback. We also would like to acknowledge Prof. Soonjo Kwon, who developed the STEP to X3D Translator (STP2X3D).

Disclaimer

The work presented in this paper 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] Egger, J., and Masood, T., 2020. “Augmented reality in support of intelligent manufacturing–a systematic literature review”. *Computers & Industrial Engineering,* ***140***, p. 106195.

[2] Pedone, G., and Mezgár, I., 2018. “Model similarity evidence and interoperability affinity in cloud-ready industry 4.0 technologies”. *Computers in Industry,* ***100***, pp. 278–286.

[3] Kwon, S., Monnier, L. V., Barbau, R., and Bernstein, W. Z., 2020. “Enriching standards-based digital thread by fusing as-designed and as-inspected data using knowledge graphs”. *Advanced Engineering Informatics,* ***46***, p. 101102.

[4] Hanke, A., Vernica, T., and Bernstein, W. Z., 2020. “Linking performance data and geospatial information of manufacturing assets through standard representations”. In ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[5] ASME Y14.5-2009, 2009. *Dimensioning and Tolerancing*. American Society of Mechanical Engineers, New York, USA.

[6] Pratt, M. J., Anderson, B. D., and Ranger, T., 2005. “Towards the standardized exchange of parameterized feature-based cad models”. *Computer-Aided Design,* ***37***(12), pp. 1251–1265.

[7] ISO 10303-242:2014, 2014. *Industrial automation systems and integration – Product data representation and exchange - Part 242: Application protocol: Managed Model-based 3D Engineering*. International Organization for Standardization, Geneva, Switzerland.

[8] Feeney, A. B., Frechette, S. P., and Srinivasan, V., 2015. “A portrait of an iso step tolerancing standard as an enabler of

[14] https://github.com/usnistgov/TrackingExpertPlus

smart manufacturing systems". *Journal of Computing and Information Science in Engineering,* ***15***(2).

[9] AP242.org, 2009. Development of a Convergent Modular STEP Application Protocol Based on AP 203 and AP 214: STEP AP 242 – Managed Model Based 3D Engineering. Version 1.0.

[10] ISO 1101:2012, 2012. *Geometrical product specifications (GPS) – Geometrical tolerancing – Tolerances of form, orientation, location, and run-out*. International Organization for Standardization, Geneva, Switzerland.

[11] Hall, S., and Takahashi, R., 2017. "Augmented and virtual reality: The promise and peril of immersive technologies". *Media & Entertainment*, October.

[12] Martin, E., 2019. Get to the point: Pixyz 2019.2, Dec.

[13] Urbas, U., Vukašinović, N., et al., 2019. "Displaying product manufacturing information in augmented reality for inspection". *Procedia CIRP,* ***81***, pp. 832–837.

[14] Kleverud, F., 2018. "Interactive visualization of cad data in real time rendering". Master's thesis.

[15] Fiorentino, M., Monno, G., and Uva, A., 2009. "Tangible digital master for product lifecycle management in augmented reality". *International Journal on Interactive Design and Manufacturing (IJIDeM),* ***3***(2), pp. 121–129.

[16] Fiorentino, M., Uva, A. E., and Monno, G., 2011. "Product manufacturing information management in interactive augmented technical drawings". In World Conference on Innovative Virtual Reality, Vol. 44328, pp. 113–122.

[17] Schroeder, W. J., Zarge, J. A., and Lorensen, W. E., 1992. "Decimation of triangle meshes". In Proceedings of the 19th annual conference on Computer graphics and interactive techniques, pp. 65–70.

[18] ISO 10303-21:2002, 2002. *Industrial automation systems and integration - Product data representation and exchange - Part 21: Implementation methods: Clear text encoding of the exchange structure*. International Organization for Standardization, Geneva, Switzerland.

[19] ISO/IEC 19775:2013, 2013. *Information technology — Computer graphics, image processing and environmental data representation — Extensible 3D (X3D).* International Organization for Standardization, Geneva, Switzerland.

[20] Stefan, L., Hermon, S., and Faka, M., 2018. "Prototyping 3d virtual learning environments with x3d-based content and visualization tools". *BRAIN. Broad Research in Artificial Intelligence and Neuroscience,* ***9***, pp. 6–20.

[21] Ishii, H., 2008. "The tangible user interface and its evolution". *Communications of the ACM,* ***51***(6), pp. 32–36.

Implementation and Evaluation of a WLAN IEEE 802.11ay Model in Network Simulator ns-3

Hany Assasa[*][†]
IMDEA Networks Institute
Madrid, Spain
hany.assasa@imdea.org

Nina Grosheva[*]
IMDEA Networks Institute and
Universidad Carlos III de Madrid
Madrid, Spain
nina.grosheva@imdea.org

Tanguy Ropitault
Associate, National Institute of
Standards and Technology
Prometheus Computing LLC
Sylva, NC, USA
tanguy.ropitault@nist.gov

Steve Blandino
National Institute of Standards and
Technology
Gaithersburg, MD, USA
steve.blandino@nist.gov

Nada Golmie
National Institute of Standards and
Technology
Gaithersburg, MD, USA
nada.golmie@nist.gov

Joerg Widmer
IMDEA Networks Institute
Madrid, Spain
joerg.widmer@imdea.org

ABSTRACT

The IEEE Task Group ay (TGay) has recently defined new physical and medium access control specifications to design the next-generation wireless standard in the 60 GHz band, the so-called IEEE 802.11ay. Build upon its 802.11ad predecessor, IEEE 802.11ay aims to offer unprecedented performance (100 Gbps throughput, ultra-low latency) by introducing various technological advancements such as multiple-input and multiple-output (MIMO) communication, channel bonding/aggregation, and new beamforming techniques. Such performance paves the way to new emerging wireless applications such as millimeter-wave distribution networks, data center inter-rack connectivity, mobile offloading, augmented reality (AR)/virtual reality (VR), and 8K video streaming. Studying and analyzing these new use-cases is of paramount importance and demands high fidelity network-level simulator due to the scarcity/costs of real IEEE 802.11ay test-beds.

In this paper, we present our implementation of the IEEE 802.11ay standard in the network simulator ns-3. Our implementation captures the specifics of IEEE 802.11ay operations such as 11ay frame structure, channel bonding, new beamforming training procedures, quasi-deterministic MIMO channel support, and Single-User (SU)-MIMO (SU-MIMO)/ Multi-User (MU)-MIMO (MU-MIMO) beamforming training. We validate and demonstrate by simulations the performance of the aforementioned techniques. The code for our simulation model is publicly available.

[*] Both authors contributed equally to this work.
[†] The author is currently affiliated with Pharrowtech.

CCS CONCEPTS

• **Networks** → **Network simulations**; **Wireless local area networks**.

KEYWORDS

Millimeter Wave, IEEE 802.11ay, 60 GHz, WiGig, MIMO, ns-3, Simulations

ACM Reference Format:
Hany Assasa, Nina Grosheva, Tanguy Ropitault, Steve Blandino, Nada Golmie, and Joerg Widmer. 2021. Implementation and Evaluation of a WLAN IEEE 802.11ay Model in Network Simulator ns-3. In *2021 Workshop on ns-3 (WNS3 2021), June 23, 2021, Virtual.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The millimeter-wave (mmWave) band has become immensely popular in the recent past. Many mobile network operators around the world started rolling out the mmWave spectrum in their 5G mobile systems to alleviate the current wireless capacity crunch. Besides, consumer-grade devices are increasingly including mmWave support. The IEEE 802.11ad standard [9], introduced in 2012, was the first wireless local area networks (WLAN) standard to provide medium access control (MAC) and physical (PHY) specifications for wireless networking in the unlicensed 60 GHz band. Despite the technical achievement that IEEE 802.11ad represented at its release (around 6.72 Gbps throughput), this standard never fully took benefit of the vast capacities of the 60 GHz band. Many emerging wireless applications such as mmWave distribution networks, uncompressed content streaming for VR/AR technologies, and dense network deployments proved to be hardly attainable with IEEE 802.11ad. The main reasons lie in the fact that first, the standard was not designed with network scalability in mind. Then, it did not exploit advanced PHY layer technologies such as MIMO and channel bonding that can boost its performance/reliability by order of magnitudes. Implementing these PHY layer technologies is challenging due to the wide communication bandwidth in the mmWave band which exacerbates linear and non-linear impairments at the

Assasa, Hany; Grosheva, Nina; Ropitault, Tanguy; Blandino, Steve; Golmie, Nada T.; Widmer, Joerg. "Implementation and Evaluation of a WLAN IEEE 802.11ay Model in Network Simulator ns-3." Presented at Workshop on ns-3 (WNS3 2021), Gaithersburg, MD, US. June 23, 2021 - June 25, 2021.

Radio Frequency (RF) devices. However, the recent advancements in the design and fabrication of mmWave electronics paved the way towards high performance, robust, low-power consumption, and low-cost radio-frequency integrated circuits (RFICs).

This motivated the WiFi alliance to form the TGay in 2015 to define the next-generation mmWave standard, the so-called IEEE 802.11ay [1]. The following design factors were taken into account during the standardization phase: i) The standard must support a throughput of at least 20 Gbps. ii) It must maintain backward compatibility with IEEE 802.11ad. iii) It must extend the set of possible use cases and scenarios by introducing novel solutions at the MAC and PHY layers. Most of these requirements are achieved thanks to the incorporation of advanced physical layer solutions that are predominant in wireless systems operating at sub-6-GHz. These solutions include MIMO, channel bonding and aggregation, fast beamforming training, and multi-user transmission. At the time of writing, no IEEE 802.11ay compliant commercial off-the-shelf (COTS) devices or network-level simulators exist which hinders research progress and innovation. In this work, we fill this gap by introducing our implementation for the IEEE 802.11ay in the popular network simulator ns-3. The main contributions of our paper are as follows:

- We upgrade our ns-3 IEEE 802.11ad model [3–5] to support IEEE 802.11ay. This includes defining 802.11ay frame structure, modulation and coding schemes (MCSs), channelization, and error-model.
- We add support for all enhanced directional multi-gigabit (EDMG) training (TRN) field variants.
- We extend our Quasi-Deterministic (Q-D) channel model to support MIMO communication.
- We introduce MIMO analog beamforming training procedure for both SU-MIMO and MU-MIMO cases. Additionally, we implement SU-MIMO channel access procedure.
- Finally, we make our implementation publicly available for the research community.

The paper is structured as follows. In Section 2, we provide background on the new technologies introduced in IEEE 802.11ay with a special focus on the differences with its predecessor IEEE 802.11ad. Section 3 presents our ns-3 IEEE 802.11ay implementation, and Section 4 highlights our evaluation campaign for the proposed model. Finally, Section 5 concludes the paper.

2 BACKGROUND ON IEEE 802.11AY

In this section, we briefly present the major new features of the PHY and MAC layers of the IEEE 802.11ay standard.

2.1 EDMG Waveform

Figure 1 depicts the EDMG frame format. To maintain backward compatibility with IEEE 802.11ad, the EDMG frame reuses both the directional multi-gigabit (DMG) Preamble and DMG Header fields. Thus, the EDMG frame is divided into two parts. The first part is referred to as the Non-EDMG portion and is recognizable by DMG devices. The second part, which is known as the EDMG portion, contains all the fields that are recognized by EDMG stations (STAs).

Similar to IEEE 802.11ad, IEEE 802.11ay supports three types of physical layers technologies: Control, Single Carrier (SC), and Orthogonal Frequency Division Multiplexing (OFDM). The Control PHY is dedicated to the transmission of management and control frames such as DMG Beacons and beamforming training frames. Thus, it is designed to be robust for communication in low Signal-to-Noise Ratio (SNR) conditions. All frames transmitted in this mode can be recognized and decoded by legacy DMG devices.

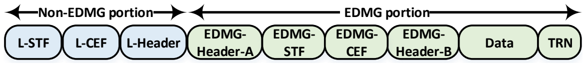


Figure 1: EDMG Waveform.

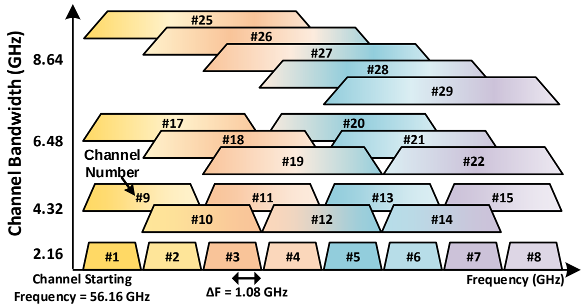


Figure 2: EDMG Channel Configurations.

For data communication, either EDMG SC or EDMG OFDM can be used. The standard mandates the support of EDMG SC mode MCSs 1 - 5 and 7 - 10 with a single spatial stream. The SC PHY has an extended set of MCSs (1 to 21) with a maximum PHY throughput of 8085 Mbps per spatial stream over a single channel for a normal guard interval (GI). On the other hand, EDMG OFDM defines 20 unique MCSs with a maximum throughput of 8316 Mbps. The support of EDMG OFDM is optional.

2.2 Channel Configuration

In IEEE 802.11ad, the 60 GHz band covered operation from 57 GHz to 64 GHz divided into four channels of 2.16 GHz. Communication at this frequency range suffers from high oxygen absorption which limits the communication range. With the growing interest in fixed wireless access (FWA) deployment and the adoption of the unlicensed mmWave band for backhauling and fronthauling, the FCC decided to double the bandwidth to cover from 57 GHz to 71 GHz, allowing a total of 14 GHz of unlicensed spectrum. The new frequency range between 64 GHz and 71 GHz does not suffer from high oxygen absorption which makes it suitable for backhauling applications where long-range communication is needed.

Figure 2 shows all the possible channel configurations for IEEE 802.11ay. IEEE 802.11ay supports operation in eight 2.16 GHz channels. To increase the data rate further, IEEE 802.11ay allows bonding a contiguous set of channels to obtain a larger channel. A maximum of four channels can be bonded which results in channel width of 8.64 GHz. The standard mandates the support of two bonded channels (4.32 GHz).

2.3 Beam Refinement Protocol

IEEE 802.11ad introduced the beam refinement protocol (BRP) to refine the beams obtained from the beamforming training (BFT) in

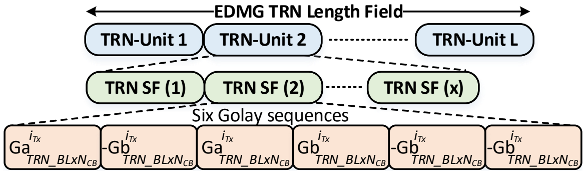

EDMG TRN Length Field
TRN-Unit 1
TRN-Unit 2
TRN-Unit L
TRN SF (1)
TRN SF (2)
TRN SF (x)
Six Golay sequences
$Ga^{i_{Tx}}_{TRN\_BLxN_{CB}}$
$-Gb^{i_{Tx}}_{TRN\_BLxN_{CB}}$
$Ga^{i_{Tx}}_{TRN\_BLxN_{CB}}$
$Gb^{i_{Tx}}_{TRN\_BLxN_{CB}}$
$-Gb^{i_{Tx}}_{TRN\_BLxN_{CB}}$
$-Gb^{i_{Tx}}_{TRN\_BLxN_{CB}}$

Figure 3: EDMG TRN Field Structure

the Sector Level Sweep (SLS) phase. The beam refinement protocol (BRP) appends a special element, called the TRN field, at the end of the packet to perform fast beam switching across multiple narrow beam patterns within the same packet. IEEE 802.11ad mandates that any signal transient that occurs due to the change of a beam pattern must settle within 36 ns. Building an RFIC with such specifications is challenging and requires an optimized analog and digital architecture. Due to these constraints, many COTS devices either omit the BRP support or implement a proprietary version with a relaxed switching time. To tackle this issue, IEEE 802.11ay performed a major redesign of the TRN field to cope with heterogeneous hardware capable end-devices.

Figure 3 shows the EDMG TRN field structure. A TRN field is composed of a variable number of TRN-Units. Each TRN unit in turn contains multiple TRN subfields where a single TRN SF contains six Golay sequences. IEEE 802.11ay introduces a variable size of the Golay sequence that can be configured by the user and additionally, in the case of channel bonding, depends on the number of continuous channels. Golay sequences have very robust correlation properties which make them suitable for channel estimation. As a result, IEEE 802.11ay defines a unique and orthogonal set of Golay sequences for each space-time stream (i_{Tx}) to facilitate channel estimation during MIMO communication.

2.4 MIMO Communication

In IEEE 802.11ad, even though a DMG STA can have multiple Phased Antenna Arrays (PAAs) connected to its RF chain, only a single PAA can be utilized at a time which results in a single stream transmission. This motivated IEEE 802.11ay to adopt MIMO support as a way to increase its throughput by multi-fold. IEEE 802.11ay supports concurrent transmission and reception of eight spatial streams at the same time and over the same frequency. The standard mandates the support of analog RF precoding for MIMO communication. In this mode, PAA can synthesize a narrow beam pattern to create multiple orthogonal spatial channels for each stream. However, depending on the quality of the phase shifters and the geometry of the PAA, generating a pencil beam pattern is not always feasible. Thus, IEEE 802.11ay proposes a hybrid analog and digital beamforming protocol to compensate for non-idealities in the analog domain and achieve the maximum gain of a MIMO system.

IEEE 802.11ay implements two variants of MIMO transmission. The first variant is known as SU-MIMO which allows transmitting and receiving multiple spatial streams (up to eight) between two devices. The second type is known as downlink MU-MIMO. In this type, an access point (AP) can transmit multiple spatial streams to multiples users (up to 8) at the same time.

3 IMPLEMENTATION

In the following section, we present the design and the implementation details of our IEEE 802.11ay model in ns-3. Our implementation is publicly available on GitHub [2].

3.1 IEEE 802.11ay Framing

As presented in section 2.1, IEEE 802.11ay introduces a new set of MCSs for both EDMG SC and EDMG OFDM with the addition of 64-QAM, a high order modulation scheme. Our implementation supports all of these new MCSs. Besides, we provide a detailed PHY layer model for transmitting and receiving different fields in the EDMG Physical Layer Convergence Protocol (PLCP) frame. To ensure accurate simulations, we integrate IEEE 802.11ay SNR to bit error rate (BER) lookup tables (LUTs) generated by NIST 802.11ay link-level simulator [10].

3.2 EDMG TRN Field

We implement the flexible and configurable TRN field structure presented in section 2.3. Additionally, we incorporate the corresponding state machines for transmitting and receiving all variants such as EDMG BRP-TX, EDMG BRP-RX, and EDMG BRP-RX/TX. Interested readers can refer to [8] for further details on the different variants.

Figure 4 shows the various states for transmitting EDMG BRP-TX and EDMG BRP-RX frames. During the transmission of EDMG BRP-RX frame, the grey blocks are omitted and M is set to 10. The EDMG BRP-RX/TX variant is used for transmit and receive beamforming training at the same time. This TRN structure is newly introduced in IEEE 802.11ay and is used for both single-input and single-output (SISO) and MIMO BFT. Due to space constraints, we show only the state-machine for transmitting EDMG BRP-TX and EDMG BRP-RX.

3.3 MIMO Q-D Channel Generation

In [5], we presented the Q-D channel model which was added to our IEEE 802.11ad implementation. The channel realizations were generated by the NIST Q-D Channel Realization Software [6], which is a full 3D ray-tracing model that captures the geometrical properties of the channel for each point-to-point pair. The software generates a 3-D multi-point to multi-point double directional channel impulse response (CIR) providing the details of the magnitude, phase, and time of arrival, direction of departure (DOD), and direction of arrival (DOA) of individual propagation paths between multiple points in space. To enable MIMO channel realization, the NIST Q-D channel Realization software has been augmented to allow the generation of the point-to-point CIR not only between each device pair, but also between each device's PAAs pair.

3.4 MIMO Operation

We extend the `QdPropagationEngine` class to include a MIMO engine that handles the calculation of the received signal power whenever a transmission is initiated with more than one active PAA. Our approach avoids the scheduling of multiple events for the different streams transmitted to guarantee the same simulation scalability as SISO. On the transmitter side, a single transmission event is scheduled and the transmit power is allocated equally

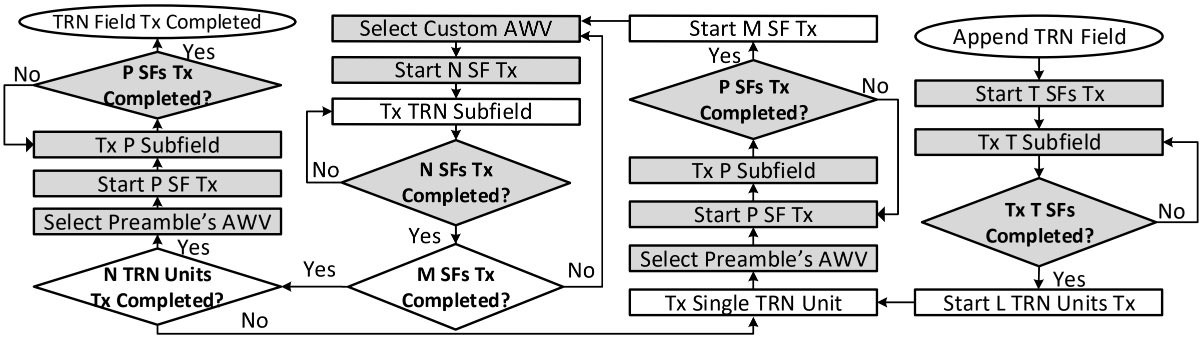


Figure 4: EDMG BRP-TX & EDMG BRP-TX Transmit State Machine Implementation.

between the transmit PAAs. On the receiver side, the MIMO engine uses the MIMO Q-D channel realizations provided by the NIST Q-D Channel realization software to calculate the received signal power for each pair of active transmit and receive PAAs. The `DmgWifiPhy` class then receives a list of RX signal powers and handles the event reception according to the type of MIMO transmission (e.g, data, beamforming training, etc.).

In the case of SU-MIMO data communication, a packet decoding operation is scheduled as explained in Section 3.6. However, for BRP packets transmitted during the MIMO BFT procedures, a different approach is necessary. The standard specifies that these packets are transmitted using spatial expansion, i.e a single space-time stream is mapped to all transmit chains active with a relative cyclic shift between the different chains. This allows the receiver to separate signals coming from the different transmit PAAs and removes unintended beamforming effects. In our implementation, the effect of spatial expansion is modeled by only attempting to decode the stream with the highest received power, considering that the cyclic shift diversity will be sufficient to remove the interference from the other received streams. The decoding of the packet afterward follows the standard SISO procedure. The TRN field of the BRP packets is also transmitted in MIMO mode and is composed of orthogonal waveforms. This orthogonal design allows us to train multiple transmit and receive antennas simultaneously by extracting the TRN-SF of each stream without any interference. Therefore, for MIMO TRN SFs, we can calculate the SNR of each received stream. These values are calculated without taking into account any inter-stream interference and are equivalent to SISO transmissions. Additionally, we add the possibility to calculate the signal-to-interference-plus-noise ratio (SINR) values of each TRN SF. These values are calculated by adding the received power from the other TX antennas active as inter-stream interference. We use the SNR values in the SISO phase of SU-MIMO BFT in order to get accurate measurements for the SISO performance, and we use the SINR later in the MIMO phase of SU-MIMO and MU-MIMO BFT to evaluate the effects of inter-stream interference.

3.5 MIMO Beamforming Training

MIMO communication involves using multiple transmit and receive PAAs to transmit data in several spatial streams. To be able to successfully establish independent streams, it is crucial to minimize the inter-stream interference to enable sufficient per-stream SINR for data decoding. To this end, IEEE 802.11ay introduced MIMO BFT. MIMO BFT is a very challenging task since an exhaustive evaluation of all the possible PAA stream configuration combinations is not viable in real-world MIMO implementations. For example, using a small codebook with 27 predefined sectors in a 2x2 MIMO setup would require testing over half a million different combinations.

IEEE 802.11ay decided to decouple MIMO BFT in two phases to overcome this problem: the SISO phase and the MIMO phase. The SISO phase aims to obtain the optimal SISO BFT for every SISO transmit/receive PAA pair of the MIMO communication. Even though these results do not provide an estimation of the inter-stream interference, they can be used to identify/select a promising subset of candidates to evaluate in the MIMO phase, enabling scalability. In the following MIMO phase, the different transmit and receive MIMO candidate combinations are tested and the MIMO performance, including the inter-stream interference effect, is measured.

The selection of candidates to test in the MIMO phase is implementation specific and not defined by IEEE 802.11ay. Thus, for the transmit training, we developed a custom approach based on [7], which suggests assigning to different beam pattern combinations a joint-beam score and to select the MIMO phase candidates from the top K combinations. In our implementation, the joint-beam score is the sum of the individual transmit beam patterns SNRs obtained in the SISO phase. The implementation can be easily extended to other selection algorithms. The list of transmit candidates given by our algorithm is trained in the MIMO phase. Each candidate is comprised of a TX configuration for each PAA involved in the MIMO training.

For the receive training, our implementation uses a different approach. This comes from our observation that the measurements at one RX PAA are independent of the configuration of the other RX PAAs. This means that instead of testing specific RX combinations, it is possible to just test each RX sector once and then, in post processing, determine the performance of different combinations by combining the measurements taken at the different PAAs. Therefore, for the receive training in the MIMO phase, we implement a simultaneous sweeping with all PAAs across all sectors. This allows us to greatly improve the scalability of the MIMO phase as the overhead of the receive training is determined by the number of predefined sectors in the codebook and does not increase with the number of PAAs being trained.

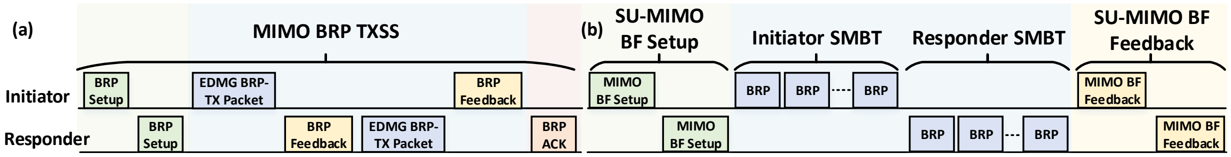


Figure 5: SU-MIMO Beamforming Training Phases: a) SISO Phase; b) MIMO Non-reciprocal Phase.

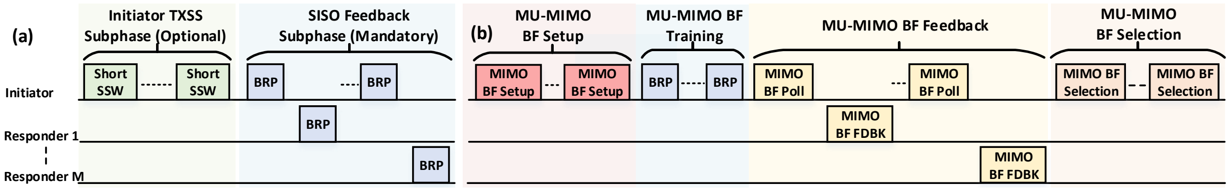


Figure 6: MU-MIMO Beamforming Training Phases: a) SISO Phase; b) MIMO Non-reciprocal Phase.

Additionally, in the MIMO phase, we implement an option to refine the beam selection by testing different antenna weight vectors (AWVs) for each sector. As the accurate estimation of the inter-stream interference is crucial to the MIMO phase, if this option is activated, all possible combinations of the transmit AWVs are tested. The number of possible combinations increases exponentially with the number of active PAAs and therefore this option improves the accuracy of the chosen beams but reduces the scalability of the MIMO phase training.

After the MIMO phase is completed, it is necessary to rank the performance of the different combinations tested and determine what is the optimal MIMO configuration. To this end, we choose the combinations that maximize the minimum per stream SINR as it will maximize the possibility that multiple spatial streams can be established.

It is important to note that in our implementation, we make no assumptions about the transmit and receive PAA pairs that establish the streams. Instead, all possible pairs are tested and the optimal combination is selected. Additionally, we added traces to allow the user to obtain the full set of SISO and MIMO phase measurements, as well as the chosen lists of TX candidates by our selection algorithm. In this way, the user can gain significant insight into the MIMO performance and evaluate the MIMO BFT algorithms.

We implement standard-compliant SU-MIMO and MU-MIMO BFT algorithms. IEEE 802.11ay specifies that the SISO feedback can be obtained from a previous SISO BFT or an optional new SISO transmit sector sweep (TXSS) can be performed. In both algorithms, we choose to support the SISO TXSS subphases to guarantee the most-up-to-date SISO feedback, as in this case the training is executed just before the MIMO phase. Additionally, the MIMO phase can be non-reciprocal or reciprocal, depending on whether the STAs involved in the training support antenna pattern reciprocity, meaning that we can consider that the transmit antenna configurations will be the same as the receive antenna configurations. For now, we support the non-reciprocal MIMO phase as it must be supported by all MIMO capable STAs and can also be used in reciprocal scenarios. Below we discuss the specifics of the SU-MIMO and MU-MIMO algorithms we implemented.

3.5.1 SU-MIMO Beamforming Training. The SU-MIMO BFT algorithm enables training between two SU-MIMO capable STAs. It includes training of the transmit and corresponding receive antenna configurations for both STAs involved, which means that after the conclusion of the BFT SU-MIMO communication can be established in both directions.

In the SISO phase, only transmit training is performed using BRP packets with transmit training (TRN-T) SFs transmitted and received with multiple active PAAs. As explained in Section 3.4, the orthogonal design of the MIMO TRN field in these packets allows us to determine the SNR values of each transmit chain without considering any inter-stream interference. In this way, multiple PAAs can be simultaneously trained which significantly reduces the training duration and increases the scalability as the number of PAAs being trained increases.

The MIMO phase, on the other hand, involves both transmit and receive training of MIMO combinations. This is done with BRP packets with TRN-R/T SFs, which enable simultaneous transmit and receive training. The same transmit configuration is kept for as many TRN Units as the Responder has requested for receive training. During the reception of these Units, the Responder switches the RX configuration at the start of each TRN SF. As we explained in Section 3.4, in this phase we record the calculated SINR values that allow us to estimate the inter-stream interference.

Figure 5 shows our SU-MIMO BFT algorithm implementation.

3.5.2 MU-MIMO Beamforming Training. The MU-MIMO protocol, shown in Figure 6, is conceptually very similar to the SU-MIMO BFT protocol described in Section 3.5.1, with two main differences. First, during the MU-MIMO BFT an Initiator trains with multiple Responders from a MU group, requiring a modification of the Feedback phases to a poll and response format. Second, IEEE 802.11ay only defines MU-MIMO transmissions in the downlink direction and performs only transmit training for the Initiator and receive training for the Responders.

Additionally, the transmit training in the SISO phase is performed with Short Sector Sweep (SSW) packets transmitted and received in SISO mode, instead of MIMO TRN-T SFs. This is because the Initiator is training with multiple Responders and it is not possible to guarantee that all of them will be able to receive the BRP packets. In order to reduce the training time, the new short SSW frames are used, instead of legacy SSW frames. The short SSW frame is a PHY layer frame and it is 6 bytes long compared to 26 bytes for the legacy SSW which results in a 31% reduction in the transmission time. We add support for these frames by enabling the transmission of WiFi packets without a MAC header.

The MIMO training is performed using TRN-R/T SFs, same as for SU-MIMO. However, it requires an additional Selection Subphase where the Initiator informs the MU group of the Responders and optimal MIMO configurations that have been selected for MU-MIMO communication. This allows the Responders to activate the correct receive configuration when MU-MIMO transmissions take place.

3.6 SU-MIMO Channel Access Procedure and Data Transmission

IEEE 802.11ay defines various methods for SU-MIMO channel access before data transmission can take place. We implement a RTS/DMG CTS mechanism where a control trailer is added to the RTS and DMG CTS frames. The control trailer contains signaling regarding the SU-MIMO configuration used for data transmission, allowing the STAs to set up the transmit and corresponding receive antenna configurations that were previously trained.

Additionally, for the data transmission, we extend the `DmgWifiMac`, `MacLow`, `DmgWifiPhy` and `InterferenceHelper` classes to support the transmission and decoding of MIMO packets. In the Interference Helper, we calculate the per stream SINR values that take into account the inter-stream interference and use this to determine the per-stream packet success rate. Analogous to the calculation of the chunk success rate, the success rate for the packet is equivalent to the multiplication of the per-stream PSRs.

4 EVALUATION

In this section, we evaluate and validate our IEEE 802.11ay implementation in ns-3. All our simulation scenarios utilize the Q-D channel model. Simulation parameters are summarized in Table 1. All the devices in the network utilize 2x8 elements Uniform Rectangular Array (URA) PAA which yields a narrow beam in the azimuth plane, and a wide beam in the elevation plane.

4.1 Achievable Throughput

In this simulation, we evaluate the maximum achievable throughput for the IEEE 802.11ay protocol for all the EDMG MCSs with various channel widths. Our scenario consists of two IEEE 802.11ay devices separated apart by one meter and have a Line-of-sight (LOS) link. We configure the two devices to use a broadside beam pattern thus ensuring a high SNR value that prevents any packet loss. To eliminate beamforming training overhead, we install `DmgAdhocWifiMac` which is an experimental MAC layer implementation that facilitates studying PHY layer features without adding the complexity of MAC layer protocol. This MAC implementation allocates the whole Beacon Interval (BI) for data transmission.

Figure 7 depicts our simulation results for both EDMG SC and EDMG OFDM PHYs. To exclude the overhead of each layer in the protocol stack, we measure the throughput at the application layer. We can observe that the maximum achievable throughput with four bonded channels is around 29.6 Gbps for EDMG SC and 31.25 Gbps for EDMG OFDM. We notice a degradation in the throughput for EDMG-MCS-17. it is worth mentioning that this might cause issues with rate adaptation algorithm (RAA) algorithms as they would expect a monotonic increase in throughput when increasing the MCS.

Table 1: Simulations Parameters

Parameter Name	Parameter Value
Application Type	OnOffApplication
Payload Size	1472 Bytes
Transport Protocol	UDP
MAC Queue Size	4000 Packets
Aggregation Type	A-MSDU and A-MPDU
A-MSDU Max. Size	7935 Bytes
A-MPDU Max. Size	4194303 Bytes
PPDU Max. Duration	2 ms
Block ACK Size	1024 Frames
MAC Protocol	CSMA/CA
Codebook Type	Parametric
Number of Transmit Sectors	27 Sectors
Sectors Azimuth Steering Angles	-80°:20°:80°
Sectors Elevation Steering Angles	-45°, 0°, 45°
A-BFT Sectors	13 Sectors
Guard Interval Size	Normal
Transmit Power	10 dBm
Rx Noise Figure	10 dB
Operating Frequency	60.48 GHz (CH2)

The throughput obtained in our simulation considers an ideal scenario where we have neither collision on the wireless medium nor packet loss. In a real network, the actual throughput will be lower due to i) the overhead imposed by different channel access periods in the BI ii) the usage of Ready-to-Send (RTS)/Clear-to-Send (CTS) handshake protocol iii) frequent link maintenance through beamforming training in Data Transmission Interval (DTI) access period. The impact of the last point depends mainly on the size of the codebook and the number of PAAs.

4.2 SU-MIMO Beamforming Training Validation

To validate our SU-MIMO implementation, our scenario is made of an AP and a STA, each one equipped with two PAAs separated by 3cm along the x-axis, deployed in a 5x10x3m room as depicted in Figure9. Each PAA is connected to a separate transmit chain which allows a maximum of two spatial streams.

Figure 8 depicts the results from the different phases of our SU-MIMO BFT algorithm between the AP (TX) and the STA (RX). The SISO phase measurements (Figure 8 (a)) show the SNR of the different transmit sectors from both TX PAAs measured at both RX PAAs. Since the PAAs separation distance is small, we can observe that the SNRs from the same transmit Sector at both receiver's PAAs are very similar in most cases. The SISO results then serve as input to our selection algorithm that selects the top K combinations as shown in Figure 8 (b). The list of K candidates is tested in the MIMO phase (Figure 8 (c)) resulting in a set of SINR measurements. For this scenario, we used top K=85 combinations tested, as we observed that this value offers a good compromise between scalability and accurate SU-MIMO configuration. In Figure 8 (d) we present a heatmap of minimum per stream SINR for each candidate tested. On the x-axis, we show the different TX candidates according to

Assasa, Hany; Grosheva, Nina; Ropitault, Tanguy; Blandino, Steve; Golmie, Nada T.; Widmer, Joerg. "Implementation and Evaluation of a WLAN IEEE 802.11ay Model in Network Simulator ns-3." Presented at Workshop on ns-3 (WNS3 2021), Gaithersburg, MD, US. June 23, 2021 - June 25, 2021.

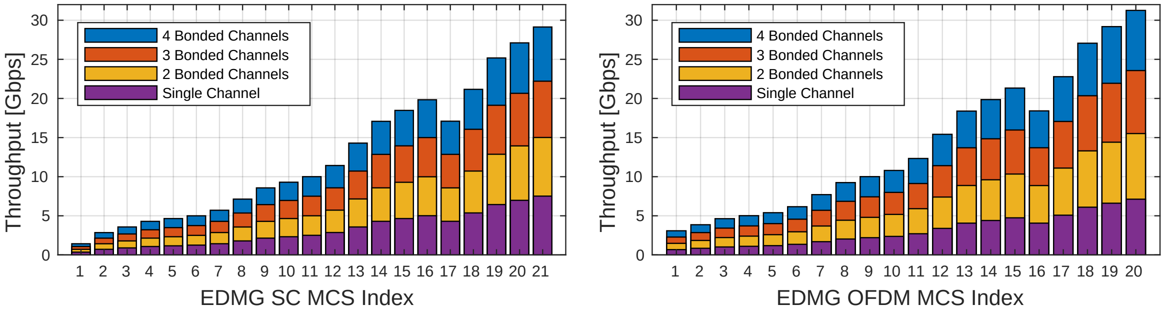


Figure 7: EDMG MCSs Throughput for Different Channel Sizes.

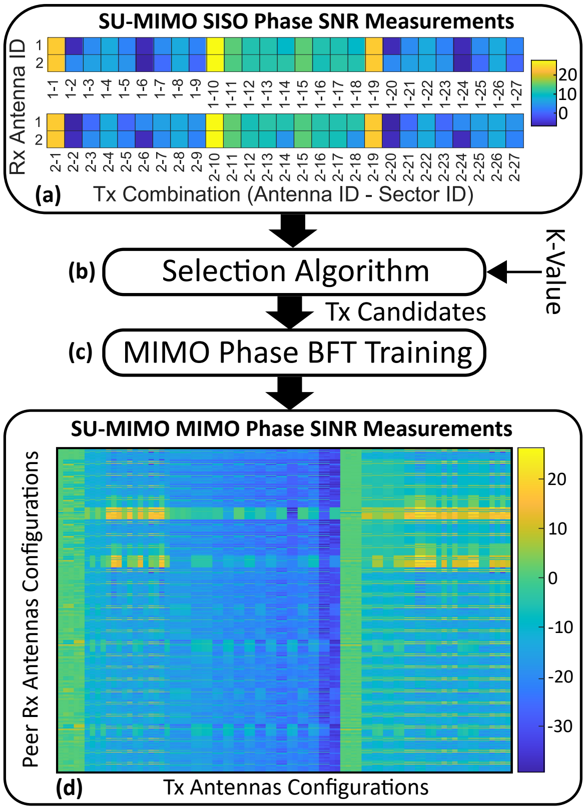


Figure 8: MIMO Beamforming Training Results.

their ranking by the selection algorithm, the first column representing the candidate with the highest joint SNR. On the y-axis, we present the different receive combinations tested. As explained in Section 3.5, we can determine the SINR for all possible receive combinations and we present them sequentially (the bottom row representing (RX PAA 1 - Sector 1, RX PAA 2 - Sector 1) and the top row representing (RX PAA 1 - Sector 27, RX PAA 2 - Sector 27). We can see that the highest-ranked candidates (leftmost columns) experience low SINR. For these candidates, both PAAs have beam patterns that utilize the LOS path as it gives the highest SISO SNR. However, when used for MIMO communication, such a combination results in high inter-stream interference due to the small PAA separation. This shows the significance of the MIMO phase, as the optimal SISO configurations can sometimes result in poor MIMO performance. Additionally, we can observe a high diversity in the SINR measurements for the different configurations tested. This implies that it can be extremely challenging to predict the MIMO performance from the SISO feedback and that the selection of good candidates for the MIMO phase is crucial to the overall functioning of the MIMO BFT algorithms. As mentioned in Section 3.5, our implementation was designed to be able to evaluate the effect of different selection algorithms and can therefore be of crucial interest to study mmWave MIMO behavior. Finally, in the top right half of the map, we can also see a high SINR area where the best SU-MIMO configuration is located.

Figure 9 shows a visualization of the best SU-MIMO configuration chosen by our BFT algorithm. We can clearly see that the first stream established, shown in Figure 9 (a), utilizes the reflections from the front and back walls and has very low gain for the LOS path and the reflections from the side-walls and the ceiling/ground. The second stream, (Figure 9 (b)), utilizes precisely those links and receives very low interference from the front and back wall reflections. The resulting combination shown in Figure 9 (c) has very high per stream SINR of 23.52 dB and 39.25 dB respectively, validating that our BFT algorithm can successfully determine good antenna configurations for MIMO communication.

Finally, after the BFT is completed, we validate our SU-MIMO data transmission implementation using the output of the MIMO Phase BFT training to setup transmit and receive antennas. The large SINR experienced by the two streams enables the use of EDMG-SC MCS-21 (8 Gbps). We observe an aggregate throughput of around 14 Gbps, validating the multi-stream transmission implementation.

4.3 MU-MIMO Beamforming Training Validation

In this scenario, we deploy an EDMG AP and two STAs in the same room as depicted in Figure 10. The AP is equipped with two RF chains where each chain is connected to a separate PAA, while the two STAs are each equipped with a single PAA. As a result, the AP

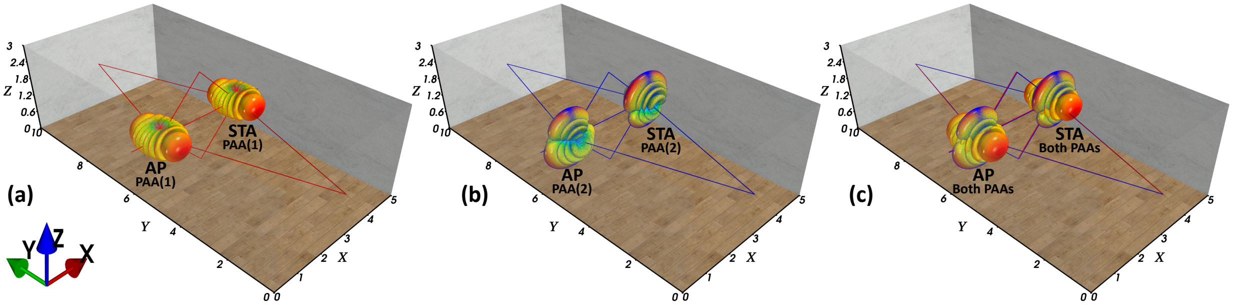


Figure 9: SU-MIMO Beamforming Training Qualitative Results: (a) PAAs Beam Patterns Corresponding to Stream 1, b)PAAs Beam Patterns Corresponding to Stream 2, (c) Combined PAAs Beam Patterns for Stream 1 and 2.

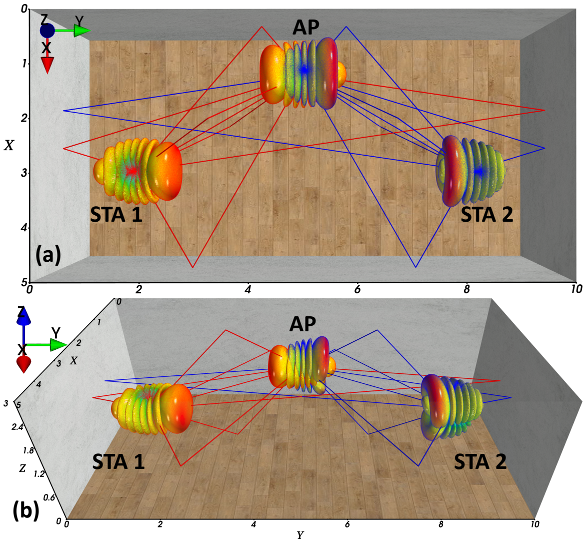


Figure 10: MU-MIMO Beamforming Training Qualitative Results (a) Top View; (b) Side View.

can transmit two spatial streams thus allowing data communication with two users at the same time.

Due to space constraints, we show only the optimal MU-MIMO configuration chosen by our algorithm in Figure 10. We can see that the high spatial separation between the STAs allows us to have two streams that utilize different multi-path components. The resulting per stream SINRs of 33.8 dB and 33.3 dB are very high and will be sufficient for MU-MIMO communication with high data rates.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we presented our implementation of the IEEE 802.11ay standard in network simulator ns-3. We implemented a diverse set of MAC and PHY features including 11ay framing, channel bonding, EDMG BRP variants, SU-MIMO beamforming training with data transmission, and MU-MIMO beamforming training. We demonstrated the maximum achievable throughput per spatial stream for each EDMG MCS for different channel configurations. Besides, we illustrated some qualitative results for SU/MU-MIMO beamforming training and beam selection algorithm.

We plan to continue improving the robustness and fidelity of our IEEE 802.11ay module. Additionally, we are working on the following features: multi-channel scheduling, MU-MIMO channel access procedure, TDD protocol, and polarization support.

ACKNOWLEDGMENTS

This work has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861222 and the National Institute of Standards and Technology (USA) under the Measurement Science and Engineering (MSE) Research Program, grant n° 60NANB19D086.

REFERENCES

[1] 2020. IEEE Draft 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 2: Enhanced Throughput for Operation in License-Exempt Bands Above 45 GHz. *IEEE P802.11ay/D7.0, December 2020* (2020).

[2] H. Assasa and N. Grosheva. 2021. *A Collection of Open-source Tools to Simulate IEEE 802.11ad/ay WLAN Networks in Network Simulator ns-3.* https://github.com/wigig-tools/ns3-802.11ad

[3] H. Assasa and J. Widmer. 2016. Implementation and Evaluation of a WLAN IEEE 802.11ad Model in ns-3. In *Proceedings of the 2016 Workshop on ns-3*. Seattle, WA, USA.

[4] H. Assasa and J. Widmer. 2017. Extending the IEEE 802.11Ad Model: Scheduled Access, Spatial Reuse, Clustering, and Relaying. In *Proceedings of the Workshop on ns-3*. Porto, Portugal.

[5] H. Assasa, J. Widmer, T. Ropitault, and N. Golmie. 2019. Enhancing the ns-3 IEEE 802.11ad Model Fidelity: Beam Codebooks, Multi-antenna Beamforming Training, and Quasi-deterministic mmWave Channel. In *Proceedings of the Workshop on ns-3*. Florence, Italy.

[6] A. Bodi, S. Blandino, J. Zhang N. Varshney, T. Ropitault, P. Testolina M. Lecci, J. Wang, C. Lai, and C. Gentile. 2021. *The NIST Q-D Channel Realization Software*. https://github.com/wigig-tools/qd-realization

[7] F. Fellhauer, N. Loghin, D. Ciochina, T. Handte, and S. ten Brink. 2017. Low complexity beamforming training method for mmWave communications. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. 1–5. https://doi.org/10.1109/SPAWC.2017.8227683

[8] Y. Ghasempour, C. R. C. M. da Silva, C. Cordeiro, and E. W. Knightly. 2017. IEEE 802.11ay: Next-Generation 60 GHz Communication for 100 Gb/s Wi-Fi. *IEEE Communications Magazine* 55, 12 (2017), 186–192. https://doi.org/10.1109/MCOM.2017.1700393

[9] IEEE. 2012. 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* (2012).

[10] N. Varshney, J. Zhang, J. Wang, A. Bodi, and N. Golmie. 2020. Link-Level Abstraction of IEEE 802.11ay based on Quasi-Deterministic Channel Model from Measurements. In *2020 IEEE 92nd Vehicle Technology Conference (VTC 2020-Fall)*. Victoria, B.C, Canada.

On the Benefits of Whole-building IAQ, Ventilation, Infiltration, and Energy Analysis Using Co-simulation between CONTAM and EnergyPlus

W. Stuart Dols[1], Chad W. Milando[2], Lisa Ng[1], Steven J. Emmerich[1] and Jyrteanna Teo[1]

1. National Institute of Standards and Technology, Gaithersburg, MD, U.S.A.
2. Boston University School of Public Health, Boston, MA, U.S.A.

8th International Building Physics Conference
Copenhagen, Denmark
25 – 27 August 2021

Abstract. Publicly available tools to perform whole-building simulation of indoor air quality, ventilation, and energy have been available for several decades. Until recently, these tools were developed in isolation, such as the whole-building contaminant transport and airflow analysis tool, CONTAM, developed by the National Institute of Standards and Technology (NIST) and the whole-building energy analysis tool, EnergyPlus, developed by the U.S. Department of Energy (DOE). The ability to couple these tools during runtime has been implemented through co-simulation, enabling improved analysis of the interdependent effects of temperature and airflow on contaminant transport and energy use on a whole-building scale.

This presentation will include the development of a set of coupled reference building models for the purposes of evaluating the potential benefits of using co-simulation between CONTAM and EnergyPlus. A set of Residential Prototype Building Models available from DOE has been modified by NIST and utilized to demonstrate the coupling process and the benefits of this coupling with respect to IAQ and energy analysis, and to evaluate multiple whole-building simulation methods related to infiltration, ventilation, and occupant exposure. These methods include an original EnergyPlus prototype model, the original model with NIST-based infiltration correlations, co-simulation between EnergyPlus and CONTAM, and stand-alone CONTAM simulations. Potential benefits will be explored related to the ability of co-simulation to address the effects of variations in building typology and ventilation system performance on contaminant transport results while leveraging the capabilities of whole-building energy analysis.

Dols, William Stuart; Milando, Chad; Ng, Lisa; Emmerich, Steven; Teo, Jyrteanna. "On the Benefits of Whole-building IAQ, Ventilation, Infiltration, and Energy Analysis Using Co-simulation between CONTAM and EnergyPlus." Presented at 8th International Building Physics Conference (IBPC 2021), Copenhagen, DK. August 25, 2021 - August 27, 2021.

1 Introduction

The multizone airflow and contaminant transport analysis software, CONTAM, has been under continuous development at the National Institute of Standards and Technology (NIST) since the 1980s [1-3]. This program is publicly and freely available for download from the NIST website. CONTAM provides the ability to simulate ventilation and indoor air quality (IAQ) on a whole-building scale; however, CONTAM does not perform heat transfer analysis. Therefore, users are required to input indoor building temperatures, which impact airflow rates. While these temperatures may be scheduled, e.g., to conform with thermostatic setpoints, they may not fully capture the heat transfer-related properties associated with building energy-related systems such as envelope construction and heating, ventilating, and air-conditioning (HVAC) systems that affect temperature differences between zones and fan run times. EnergyPlus is also publicly available software developed by the U.S. Department of Energy [4]. EnergyPlus provides the ability to simulate multizone, whole-building heat transfer analysis including the sizing and control of HVAC systems.

These programs are able to address a broad range of important processes but when used alone the capabilities are limited. For example, CONTAM has been used to evaluate the energy costs of infiltration without the direct benefit of energy simulation [5]. EnergyPlus, on the other hand, has been incorporated with the ability to simulate two contaminants: carbon dioxide (CO_2) and a generic contaminant but does not account for particle losses associated with filtration or building envelope penetration. Further, EnergyPlus can implement an Airflow Network model (AFN) to enable multizone airflow analysis. However, these capabilities are largely based on those implemented by CONTAM and its predecessors [6, 7] and can be cumbersome to implement without a graphical user interface as provided with CONTAM. Used together these programs can better capture the often-interdependent transport mechanisms, providing more comprehensive analyses of measures aimed at improving energy and IAQ performance.

1.1 Co-simulation between CONTAM and EnergyPlus

To capture the inter-dependency between temperature and airflow (and hence contaminant transport), CONTAM has been coupled with EnergyPlus [8]. The coupling between EnergyPlus and CONTAM is achieved using quasi-dynamic coupling via the Functional Mock-up Interface for Co-simulation specification as implemented in EnergyPlus [9, 10]. This method of coupling provides for the run-time exchange of data between two separate simulation programs at regular time intervals during co-simulation.

During the co-simulation EnergyPlus acts as the controlling program. Prior to transient simulations for up to one year, EnergyPlus first performs system sizing and then performs warm-up simulations whereby co-simulation occurs repeatedly over a 24-hour period until zone temperatures stabilize. During the warm-up, reversible source-sinks of CONTAM (i.e., deposition-resuspension surfaces and diffusion-based materials) can also be loaded with contaminant via the CONTAM restart file. The data exchanged during co-simulation is outlined below, and details are provided in references [2, 8].

From EnergyPlus to CONTAM

- **Zone Temperatures** and **Relative Humidity**
- **Ventilation system airflow rates** for zone supply and return airflows
- **Outdoor airflow fractions** of outdoor airflow controllers
- **Exhaust fan airflow rates**
- **Outdoor environmental data** including temperature, barometric pressure, wind speed, and wind direction
- **Output variables** user-selected from available EnergyPlus output variables

From CONTAM to EnergyPlus

- **Zone infiltration airflows**
- **Inter-zone airflows**
- **Controls values** user-defined to be exposed via the CONTAM controls network, e.g., a signal calling for ventilation airflow due to an elevated contaminant level

1.2 Residential Building Models

A multi-family building model was selected from a set of prototype building models that were originally developed in EnergyPlus for DOE by the Pacific Northwest National Laboratory (PNNL). This model was used to demonstrate the co-simulation process and to compare the capabilities of and among various simulation tools [11]. These prototype models were intended to inform the decision-making process related to developing building energy codes, i.e., International Energy Conservation Code (IECC), and they have evolved over the years along with the relevant codes and standards [12]. The EnergyPlus model used in this study consisted of slab-on-grade construction with each apartment having a forced-air HVAC system with electric resistance heating and direct-expansion cooling coils and constant exhaust ventilation.

The building model, shown in Figure 1, is a three-story, garden style apartment building with no enclosed stairwells or shafts, and each apartment modeled as a single zone 12.19 m x 9.14 m x 2.59 m high (40 ft x 30 ft x 8.5 ft). Simulations were performed using the IECC 2006 building representation for climate zone 5A (*USA_MA_Boston-Logan.Intl.AP.725090_TMY3.epw*).

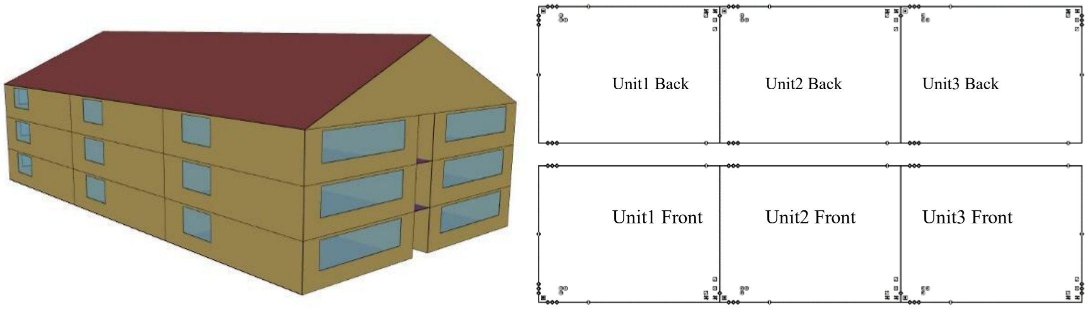


Figure 1. Multi-family building model (left: EnergyPlus model) and floor plan (right: CONTAM model) with apartment names

2 Methods

EnergyPlus version 9.1 and CONTAM 3.4 were used for the simulations, so the original EnergyPlus models were updated using the *IDFVersionUpdater* tool provided with EnergyPlus. CONTAM models were developed with the *ContamW* graphical user interface using the pseudo-geometry mode to create scaled representations of the building floor plans for each level of the building including the attic.

2.1 Coupling Strategy

Coupling of EnergyPlus and CONTAM requires building representations for both programs, i.e., an EnergyPlus input file (IDF) and a CONTAM project file (PRJ). Two NIST-developed tools were utilized to facilitate the EnergyPlus-CONTAM co-simulation: *Contam3DExport* program and *ContamFMU* dynamic link library. *Contam3DExport* creates IDF files from a PRJ file along with the files required to coordinate the co-simulation, and *ContamFMU* provides for control of software execution and the exchange of data with the CONTAM simulation engine, *ContamX*, during co-simulation. A schematic of the coupling process and associated files and software is shown in Figure 2, and details are provided in the CONTAM User Guide [2].

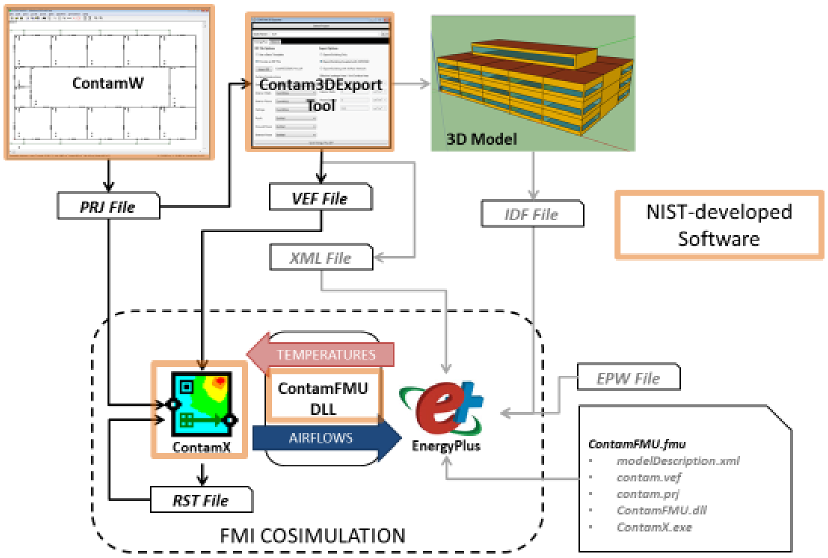


Figure 2. EnergyPlus-CONTAM Coupling Schematic

A scaled version of the building was developed in *ContamW* and extruded to a three-dimensional IDF using *Contam3DExport*. The roof as generated by *Contam3DExport* was cuboidal and required modification to create the gable geometry within the IDF. The model developed herein was within approximately one percent of the building volume of the original IDF. The exported IDF was then modified to align with the original IDF with respect to the following building model properties: internal loads, HVAC system properties (e.g., heating and cooling coils, sizing parameters, and thermostats), demand hot water, schedules, building construction and materials, and shading surfaces. Some of these items were created by *Contam3DExport* and required modification, e.g., cooling and heating coils, and other components were not included in the exported IDF file, e.g., shading surfaces.

2.2 Building Simulations

Inter-model comparisons of the resultant whole-building energy use, infiltration rates, and contaminant concentrations were made using the following simulation methods.

- Original EnergyPlus-only (labeled ***E+*** in results)
- Original EnergyPlus-only with infiltration correlations (labeled ***E+**** in results)
- EnergyPlus-CONTAM co-simulation (labeled ***COSIM*** in results)
- CONTAM-only (labeled ***Cw*** in results)

Explanations for each of these methods and relevant differences between them are provided in the following sections.

2.2.1 Energy Inputs

All the models that utilized EnergyPlus (*E+*, *E+**, and *COSIM*) implemented auto-sizing of the HVAC systems, which determined the supply fan flow rates. The CONTAM-only model (*Cw*) was defined to implement indoor temperatures based on the thermostatic set-points for both heating and cooling modes as per the EnergyPlus thermostatic set-point schedules. These schedules included the cooling season from May 1 through September 30 with a cooling setpoint of 23.88 ºC (75 ºF) and a heating setpoint of 22.22 ºC (72 ºF). Further, the supply airflow rates and an intermittent fan run time schedule (5 minutes ON, 15 minutes OFF) of the *Cw* model were based on the system sizing results and the approximate fan run-time fraction of the *E+* simulation results, respectively. In this manner, the *E+* simulation was utilized to inform inputs to the *Cw* model without direct coupling between them.

2.2.2 Infiltration Inputs

The *E+* model implemented the *ZoneInfiltration:EffectiveLeakageArea* calculation method based on Equation (1) with effective leakage areas (A_L) in cm^2, at a 4 Pa reference pressure and discharge coefficient of 1.0, for each apartment apportioned as shown in Table 1. ΔT is the indoor-outdoor temperature difference in ºC, W_s is the wind speed in m/s, and C_s and C_{ws} are stack and wind speed coefficients: 0.00029 and 0.000231, respectively. The effective leakage area in the CONTAM models (*Cw* and *COSIM*) were calculated based on the values in Table 1 and the surface areas of the *Middle* and *Top* floor apartments to be 1.443 cm^2/m^2 of wall surface area.

$$\text{Infiltration} = \frac{A_L}{1000}\sqrt{C_s\Delta T + C_{ws}\cdot W_s^2} \qquad (1)$$

Table 1. Effective Leakage Area of Original EnergyPlus Models

Apartments	*Effective Leakage Area [cm^2]*
*Corner apartments (Units 1 & 3) on Bottom and Top floor*s	286
*Center apartments (Unit 2) on Bottom and Top floor*s	252
Corner apartments (Units 1 & 3) on Middle floor	125
Center apartments (Unit 2) on Middle floor	91

The *E+** model was based on the *E+* model which was then modified to use the infiltration calculation method (referred to in EnergyPlus as *ZoneInfiltration:DesignFlowRate*) presented in Equation (2). The *Cw* model was used to determine whole-building infiltration rates for the entire year for Boston weather. The coefficients *A*, *B,* and *D* in Equation (2) were generated from these *Cw* infiltration results using the method presented in references [13, 14] and determined to be 0.4688, 0.0166, and 0.0174, respectively. The design infiltration rate, I_{design}, was set to 3.72 x 10^{-4} m^3/s per m^2 of exterior building surface area.

$$\text{Infiltration} = I_{design}(A + B|\Delta T| + D\cdot W_s^2) \qquad (2)$$

Thus, the *E+** model utilized results of CONTAM simulations to inform required inputs without direct coupling between EnergyPlus and CONTAM.

2.2.3 Ventilation Inputs

The *E+* model utilized the EnergyPlus *ZoneVentilation:DesignFlowRate* method to account for exhaust ventilation in each zone. This ventilation method acts in an additive manner with respect to infiltration to increase outdoor air intake beyond that due to infiltration. This is an empirical method of introducing outdoor air into the building as opposed to the physics-based methods incorporated by multizone or network airflow modeling. The simulations that utilized CONTAM (*Cw* and *COSIM*) incorporated exhaust ventilation via the CONTAM model. This method of incorporating exhaust ventilation acts as a driving force for infiltration as opposed to the additive nature employed by the *E+* models. The EnergyPlus correlation models (*E+**) did not implement the *ZoneVentilation:DesignFlowRate* method, because the correlations were performed with the exhaust systems activated in the CONTAM models.

2.2.4 Contaminant Inputs

To demonstrate contaminant analysis methods, two contaminants were considered: CO_2 and fine particulate matter ($PM_{2.5}$). CO_2 is one of the contaminants that EnergyPlus can simulate directly and is primarily generated by building occupants, thus it is impacted by the building ventilation rate and can be used for demand-controlled ventilation. The outdoor CO_2 concentration was set constant at 731.8 mg/m^3 (400 ppm), and the maximum CO_2 generation rate was set to 4.48 x 10^{-6} m^3/s·person which was based on the activity schedule for two occupants as defined in the original EnergyPlus model with a maximum internal heat gain of 117.28 W/person and assuming an occupant CO_2 emission rate of 3.82 x 10^{-8} m^3/s·W. $PM_{2.5}$ is associated with both internal and external sources. In these simulations, an outdoor $PM_{2.5}$ contaminant time history file was incorporated based on measurements in Boston, MA; an indoor cooking source of 1.56 mg/min was scheduled from 7:00 to 7:10 and 18:00 to 18:20 every day, and particle deposition occurred in every zone at a rate of 0.19 h^{-1} [15]. CONTAM enables filter models to be used in any airflow path including those associated with HVAC systems, i.e., outdoor air and recirculation air filters, and those associated with envelope openings, i.e., to account for particle removal as they penetrate into the building from outdoors. EnergyPlus does not enable the use of filters. Therefore, only the CONTAM models incorporated recirculation filters within the HVAC systems and an envelope penetration coefficient of 0.72 for $PM_{2.5}$.

3 Results and Discussion

Energy simulation results were evaluated to ensure the coupled model yielded reasonable results compared to the original EnergyPlus model. Comparisons between whole-building and apartment-level infiltration rates are then presented followed by contaminant results.

3.1 Energy

Annual energy usage results are shown in Figure 3 including heating, cooling, fan, and total energy use for the three simulation methods that utilized EnergyPlus (*E+*, *E+**, and *COSIM*). Results indicate that the total energy use for the *COSIM* and *E+** simulations were 12 % and 13 % less than the original results. The difference was due to the means by which exhaust ventilation was implemented in the *E+* model as will be discussed in the next section. As a result of the method used to account for exhaust in the *E+* models, consideration should be given to modifying the PNNL models to better account for infiltration.

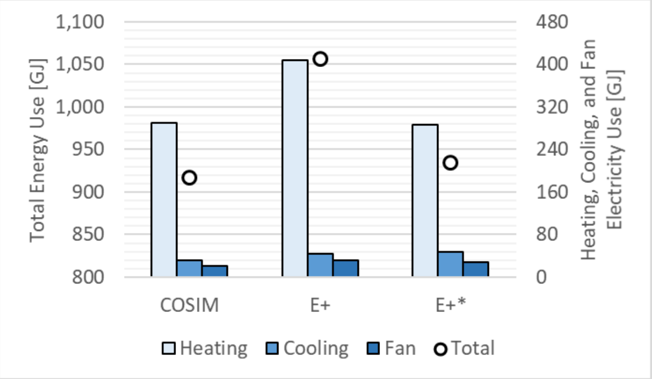


Figure 3. Annual Energy Usage Using Three Different Infiltration Calculation Methods

3.2 Infiltration

Box-whisker plots of the daily average air infiltration rates are presented in Figure 4 for the months of January and July using all four simulation methods. The *E+* results were noticeably higher than those of the other methods due to the additive nature of the exhaust ventilation and infiltration methods implemented in this model. In contrast to the "additive" nature of infiltration and exhaust flow rates in the *E+* model, the exhaust systems in the CONTAM models (*Cw* and *COSIM*) act as a driving force for infiltration, so infiltration will only be greater than the exhaust flow rate if the mass flow rate attributed to natural driving forces are greater than the exhaust ventilation rate. The *E+** results accounted for the exhaust ventilation within the correlations, i.e., the exhaust systems were active in the *Cw* simulations used to generate the correlation coefficients, so additive ventilation was not included in the IDF of the *E+** model.

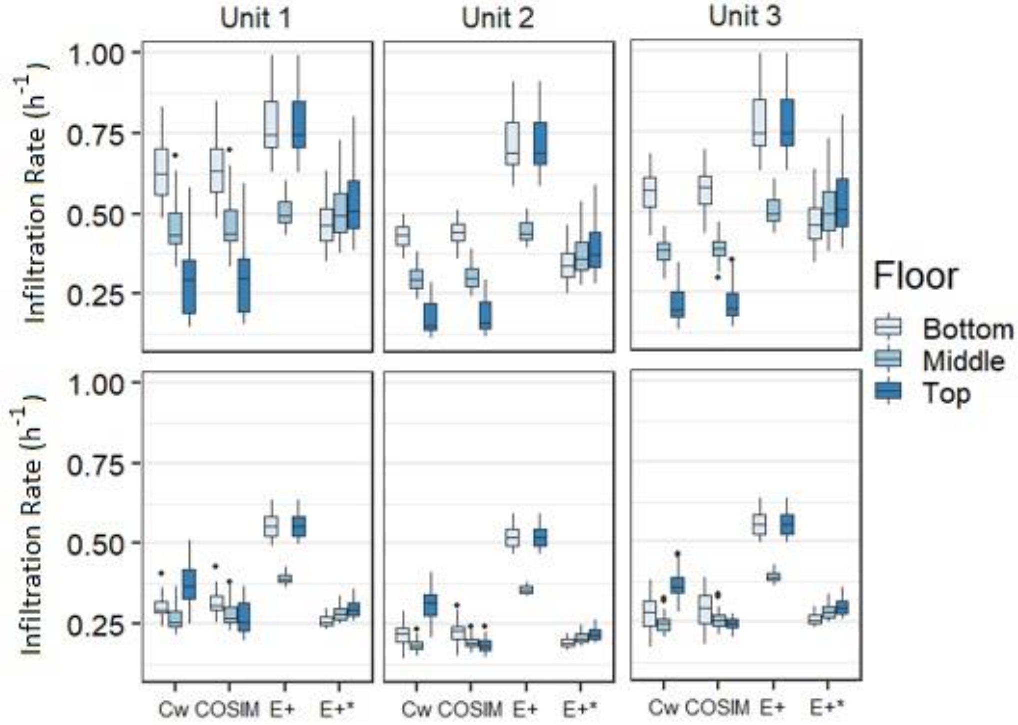


Figure 4. Daily Average Infiltration Rates of *Front* Units for January (top) and July (bottom) using Four Simulation Methods (see Figure 1 and Section 2.2 for explanations of category and zone names)

The differences in the *E+* results for different levels directly reflected the leakage areas as defined in Table 1, i.e., rates for units on the *Bottom* and *Top* floors were the same and higher than those on the *Middle* floor, and the *Front* and *Back* (not shown) units were always the same. The CONTAM-only (*Cw*) and co-simulation (*COSIM*) results were quite similar to each other (minor differences were reflected in the outliers), because the energy model controlled the single-zone temperatures very precisely to match the thermostatic set-points which were also scheduled in the *Cw* models. This reveals the capabilities of CONTAM to predict building infiltration when indoor temperatures are tightly controlled and single-zone representations are warranted. Further, the CONTAM-based calculations are not empirical, i.e., they are physics-based, so they account for pressure-driven airflows and relative leakage areas between zones including inter-floor leakages which is a key benefit of multizone network analysis.

All the simulation methods revealed the expected seasonal differences in infiltration rates, i.e., winter infiltration rates were higher than the summer rates. This difference can be attributed to greater absolute indoor-outdoor temperature differences in January than in July and associated buoyancy effects. Patterns in the relative differences between floors were the same between January (winter) and July (summer) except for the *Cw* results. Infiltration results for the simulation methods that utilized EnergyPlus (*E+*, *E+**, and *COSIM*) were obtained from the EnergyPlus output files while the *Cw* results were obtained from the CONTAM output files, and the CONTAM results included infiltration from the attic that was not accounted for in the EnergyPlus results. Therefore, in the July results, when infiltration is likely to occur from the attic into the *Top* floor zones, the infiltration rates tended to increase in the *Top* floor zones. This is another benefit of using the multizone network analysis. However, it is important to understand the meaning of simulation outputs when evaluating results, for example, when EnergyPlus infiltration rates for the *COSIM* case did not match those determined by the associated CONTAM model. The EnergyPlus simulations with correlations (*E+**) were similar in magnitude with the *Cw* and *COSIM* results, but exhibited infiltration rates that increased from the *Bottom* to the *Top* floor. This is a result of increased wind speed with elevation that is accounted for in EnergyPlus by default and the fact that the *ZoneInfiltration:DesignFlowRate* equation includes a coefficient of the square of the wind speed which dominates the resultant infiltration as presented in Equation (2).

3.3 Contaminant Transport

Contaminant results are shown in Figure 5 for the months of January and July using the four different simulation methods. Box-whisker plots of daily averages are provided for the *Front* units (results for *Back* units were very similar) and the infiltration results for *Unit 1* are repeated here to simplify evaluation of contaminant results as they relate to infiltration.

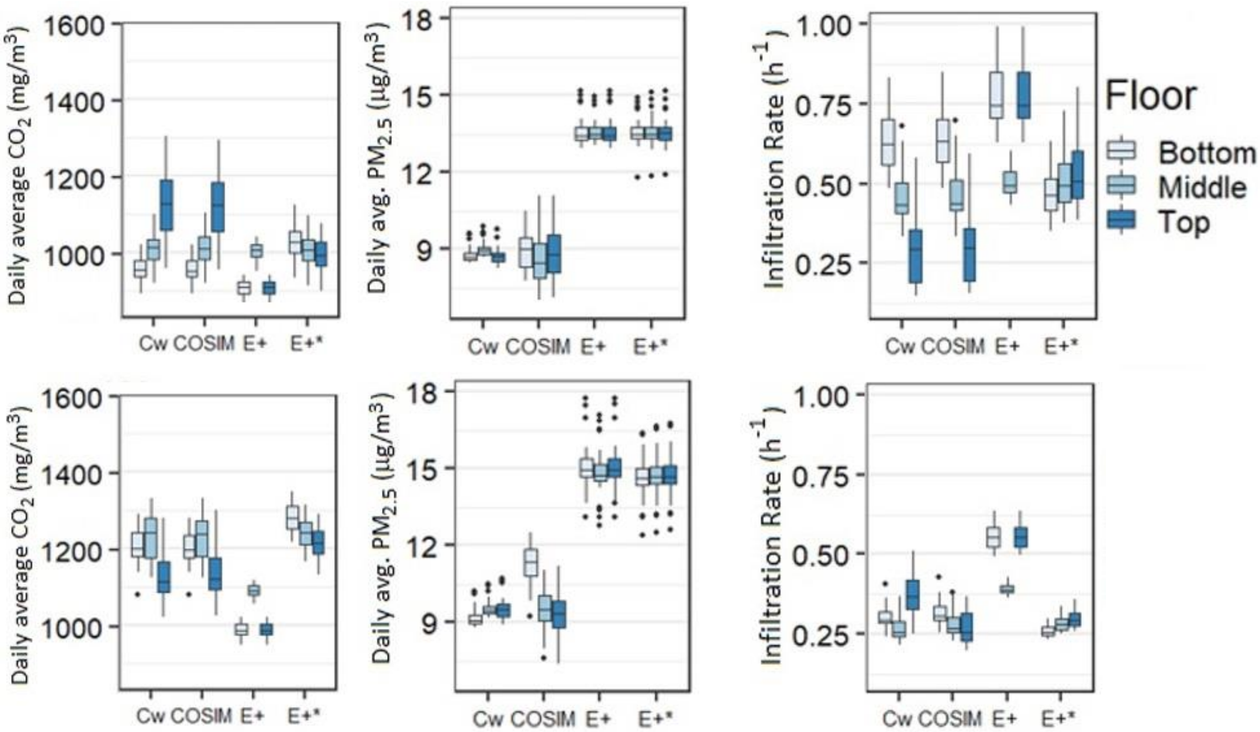


Figure 5. Daily Average Simulation Results (CO$_2$, PM$_{2.5}$, and Infiltration) of the *Front* in Unit 1 for January (top) and July (bottom) using Four Simulation Methods (see Figure 1 and section 2.2 for explanations of category and unit names).

Carbon Dioxide (CO$_2$)

CO$_2$ results generally reflected the infiltration rates because the CO$_2$ was generated internally, with higher infiltration rates leading to lower CO$_2$ concentrations. In nearly all cases, the original *E+* CO$_2$ results were lower, which follows from the additive modeling of ventilation and infiltration of the *E+* discussed previously. The variation in CO$_2$ concentrations with elevation as captured by the CONTAM methods (*Cw* and *COSIM*) also revealed the benefits of utilizing the physics-based infiltration and ventilation calculations of multizone modeling. This is revealed in *Cw* results wherein the pattern of infiltration rates is reflected in the CO$_2$ results for the July simulations. As addressed in the infiltration results, the interzone mixing between the attic and the *Top* floor enables CONTAM to capture the contaminant transport due to buoyancy-induced flows which lead to the downward internal airflows when indoor temperatures are lower than outdoors.

Particles ($PM_{2.5}$)

The average outdoor $PM_{2.5}$ concentrations (not shown) for January and July were 13.5 μg/m^3 and 16.4 μg/m^3, respectively. In the *E+* and *E+** models, the dominant removal mechanism was dilution by infiltration, so the average indoor $PM_{2.5}$ concentrations were close to the average ambient concentrations. Detailed plots of $PM_{2.5}$ concentrations for all simulation methods (not shown) revealed that concentrations fell below the ambient concentrations after each cooking event due to the combination of dilution, deposition, and filtration which was also reflected in the daily averages being below the respective outdoor averages. $PM_{2.5}$ results revealed the benefits of using CONTAM for particle analysis. The most apparent differences were the reduced levels of $PM_{2.5}$ concentrations exhibited by the *Cw* and *COSIM* results. While both EnergyPlus and CONTAM accounted for particle removal by deposition, the CONTAM models additionally removed particles via mechanical system filters and envelope penetration coefficients. All the EnergyPlus HVAC systems were auto-sized, leading to variations in system fan flows. For the *Cw* and *COSIM* cases this affected the amount of air moving across the particle filters located within the HVAC returns. This was revealed in the differences between the *COSIM* and *Cw* cases. *Cw* incorporated the fan flow rates of the *E+* results, which were different from the *COSIM* flows in most cases and significantly different in some cases, e.g., *COSIM* flows were lower than the *Cw* flows for *Unit 1* on the *Bottom* floor as shown in Figure 6.

As was the case with the previously presented infiltration rates, both the CO_2 and $PM_{2.5}$ results revealed that the contaminant concentrations in the *Front* and *Back* (not shown) units were nearly identical. This was due to the fact that these two rows of apartment buildings acted as two separate, but similar buildings. Minor differences were exhibited in the *Cw* and *COSIM* results because wind pressure coefficients on the breezeway-facing building surfaces were defined to be lower to account for shielding effects. This detailed treatment of wind pressure variations is another benefit of using multizone airflow modeling.

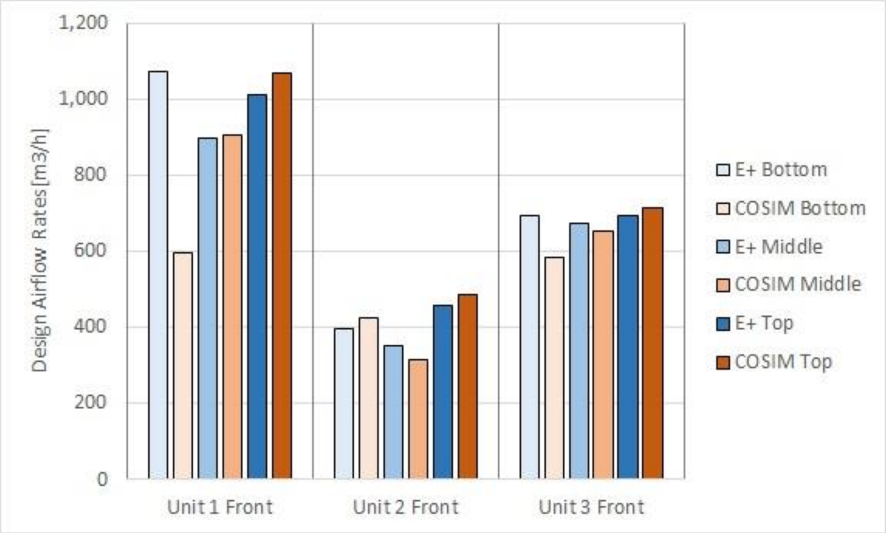


Figure 6. HVAC System Fan Flow Rates of *Front* Units for *E+* and *COSIM* Cases

4 Conclusions

A CONTAM representation of a multi-family residential reference building model was developed and coupled with EnergyPlus. This presentation provided a preliminary inter-model comparison between four different simulation methods including the original EnergyPlus model and fully coupled co-simulation between EnergyPlus and CONTAM. While some differences were revealed between these simulation methods, the benefits of co-simulation depend on the analysis being performed. In the case of these fairly simple, single-zone apartment units, major benefits of utilizing co-simulation were revealed in analyzing the removal of particulates by filtration and envelope penetration. Further, these building models utilized balanced supply and return airflows and exhaust ventilation. The balanced airflows do not drive infiltration or interzone airflow, hence contaminant transport, between apartment units, and the exhaust ventilation drove infiltration, thus reducing the effects of wind and buoyancy.

The coupled models have been shown to better capture overall building energy performance. Future work will address the effects of unbalanced system flows, variations in inter-apartment source strengths, multizone representations of apartment units, and ideally, comparison of model predictions with empirical data.

References

[1] Axley, J.W., *Indoor Air Quality Modeling: Phase II Report*, NBSIR 87-3661. 1987, National Institute of Standards and Technology: Gaithersburg, USA.

[2] Dols, W.S. and B.J. Polidoro, *CONTAM User Guide and Program Documentation Version 3.4*, NIST Technical Note 1887 Revision 1. 2020, National Institute of Standards and Technology.

[3] Walton, G.N., *AIRNET - A Computer Program for Building Airflow Network Modeling*, NISTIR 89-4072. 1989, National Institute of Standards and Technology: Gaithersburg, MD.

[4] Crawley, D.B., L.K. Lawrie, F.C. Winkelmann, W.F. Buhl, Y.J. Huang, C.O. Pedersen, R.K. Strand, R.J. Liesen, D.E. Fisher, M.J. Witte, and J. Glazer, *EnergyPlus: creating a new-generation building energy simulation program.* Energy and Buildings, 2001. **33**(4): p. 319-331.

[5] Emmerich, S.J. and A.K. Persily. *Energy Impacts of Infiltration and Ventilation in U.S. Office Buildings Using Multizone Airflow Simulation*. in *IAQ Conference*. 1998.

[6] DOE, U.S., *EnergyPlus Input Output Reference (version 9.1)*. 2019, University of Illinois: Champaign, IL.

[7] DOE, U.S., *EnergyPlus Engineering Reference (version 9.1)*. 2019, University of Illinois: Champaign, IL.

[8] Dols, W.S., S.J. Emmerich, and B.J. Polidoro, *Coupling the Multizone Airflow and Contaminant Transport Software CONTAM with EnergyPlus using Co-simulation.* Building Simulation, 2016. **9**: p. 469-479.

[9] MODELISAR, *Functional Mock-up Interface for Co-Simulation*, 2010, MODELISAR consortium.

[10] Nouidui, T., M. Wetter, and W. Zuo, *Functional mock-up unit for co-simulation import in EnergyPlus.* Journal of Building Performance Simulation, 2013. **7**(3): p. 192-202.

[11] DOE, U.S. *Building Energy Codes Program - Residential Prototype Building Models*. [cited 2021 February 4]; Available from: https://www.energycodes.gov/development/residential/iecc_models.

[12] Taylor, Z.T., V.V. Mendon, and N. Fernandez, *Methodology for Evaluating Cost Effectiveness of Residential Energy Code Changes*, PNNL-21294 Rev 1. 2015, Pacific Northwest National Laboratory: Richland, WA.

[13] Ng, L.C., N. Ojeda Quiles, W.S. Dols, and S.J. Emmerich, *Weather correlations to calculate infiltration rates for U. S. commercial building energy models.* Building and Environment, 2018. **127**(Supplement C): p. 47-57.

[14] Ng, L.C., A.K. Persily, and S.J. Emmerich, *Improving infiltration modeling in commercial building energy models.* Energy and Buildings, 2015. **88**: p. 316-323.

[15] Underhill, L.J., W.S. Dols, S.K. Lee, M.P. Fabian, and J.I. Levy, *Quantifying the impact of housing interventions on indoor air quality and energy consumption using coupled simulation models.* Journal of Exposure Science & Environmental Epidemiology, 2020. **30**(3): p. 436-447.

Dols, William Stuart; Milando, Chad; Ng, Lisa; Emmerich, Steven; Teo, Jyrteanna. "On the Benefits of Whole-building IAQ, Ventilation, Infiltration, and Energy Analysis Using Co-simulation between CONTAM and EnergyPlus." Presented at 8th International Building Physics Conference (IBPC 2021), Copenhagen, DK. August 25, 2021 - August 27, 2021.

Science and Technology Infusion Climate Bulletin
NOAA's National Weather Service
45th NOAA Annual Climate Diagnostics and Prediction Workshop
Virtual Online, 20-22 October 2020

Compound Flooding in Eastern North Carolina: Understanding Stakeholder Perceptions and Needs

Scott Curtis,[1] Jamie Kruse,[2] Anuradha Mukherji,[3] Jennifer Helgeson,[4] Kelley DePolt,[3]
Philip Van Wagoner,[3] and Ausmita Ghosh[2]

[1]Lt Col James B. Near, Jr., USAF, '77 Center for Climate Studies, The Citadel, Charleston, SC
[2]Department of Economics, East Carolina University, Greenville, NC
[3]Department of Geography, Planning, and Environment, East Carolina University, Greenville, NC
[4]National Institute of Standards and Technology, Gaithersburg, MD

ABSTRACT

While our scientific understanding of compound flood risk has made great strides in recent years, there is a lack of studies related to stakeholder awareness of the non-linear combination of pluvial, fluvial, and tidal flooding, which often occur in coastal storm environments. Here we present the concept of our NOAA-funded project "Preparing for, Responding to, and Mitigating Coastal Compound Water Hazards for Resilient Rural Communities" and describe some preliminary survey and focus group data collected from planners, emergency managers, and elected officials from across eastern North Carolina.

1. Background

The fact that Hurricanes Floyd, Matthew and Florence devastated eastern North Carolina within a period of twenty years calls for a paradigm shift in hazard preparation, response and mitigation. A common question following a storm is: "Why did my house/business flood?" Some people rely on the fact that their properties are outside the 100-year flood zone, but understanding flood risk goes beyond reliance on one tool or map. Even multiple flood risk tools that are not properly integrated can be inadequate for effective disaster management.

The hurricane hazard is composed of several storm related hazards, with water hazards: surge, pluvial flooding (flooding caused by storm water runoff), fluvial flooding, and water-borne health risks often receiving highest priority in the coastal plain of North Carolina. However, the consideration of one hazard at a time ignores how these water hazards intersect spatially and temporally. Water hazards in the storm environment are not independent of each other. For example, copious precipitation, which leads to flash flooding locally, accumulates over watersheds and is correlated to fluvial flooding. Strong storm surge, which has been related to the co-occurrence of heavy precipitation (Wahl *et al.* 2015), can also back-up riverine flow, exacerbating coastal flooding.

The combination of multiple hazards that contribute to societal, environmental or health risk is known as a compound event (Zscheischler *et al.* 2018). While compound events have been described in the climate literature, they have not been integrated into the disaster management cycle. However, these impactful events can "provide a bridge between climate scientists, engineers, social scientists, impact modelers and decision-makers, who need to work closely together to understand these complex events" (Zscheischler *et al.* 2018).

Risks, vulnerabilities and pathways to resilience in rural regions are less well studied and understood as compared to their urban counterparts (Cheng, Ganapati and Ganapati~~,~~ 2015), and rural communities tend to be disproportionately affected by compound coastal water events (CCWE) and this cumulative effect of CCWE is rarely analyzed. Economic drivers in rural communities, especially in North Carolina tend to be land- and place-based (MDC 2016); thus, the main source of economic benefit is highly sensitive to CCWE. This project focuses

Correspondence to: Scott Curtis, Lt Col James B. Near, Jr., USAF, '77 Center for Climate Studies, The Citadel, 171 Moultrie Street, Charleston, SC 29409; E-mail: wcurtis1@citadel.edu

Curtis, Scott; Kruse, Jamie; Mukherji, Anuradha; Helgeson, Jennifer; DePolt, Kelley; Van Wagoner, Philip; Ghosh, Ausmita. "Compound Flooding in Eastern North Carolina: Understanding Stakeholder Perceptions and Needs." Presented at 45th NOAA Climate Diagnostics and Prediction Workshop (CDPW). October 20, 2020 - October 22, 2020.

on rural counties in eastern North Carolina located along the coast and those adjacent to it that share estuarine environments or linked riverine systems.

2. Methods

The objectives of our NOAA-funded project "Preparing for, Responding to, and Mitigating Coastal Compound Water Hazards for Resilient Rural Communities" are to 1) assess the perceived risks and needs of the hazards management and planning community in eastern North Carolina through two-way communication, 2) examine the physical nature and economic and health impacts of CCWE from 2010 to present, and 3) use the information obtained to co-produce knowledge and tools with our study group for better preparation, response and mitigation plans. This paper focuses on objective 1 by analyzing select anonymous survey and transcript data collected before, during, and after our February 26, 2020 workshop. At this event 41 planners, emergency managers, and elected officials from across eastern North Carolina met at East Carolina University to discuss CCWE issues in small focus groups. Tabletop conversations focused on past experiences with the frequency and intensity of rain, river, and ocean induced flooding, and whether they have seen changes in the forecasting and communication of these disruptive events. Each table had a facilitator, who was a project team member or Ph.D. student, to guide discussions and a recorder, who was a student, to write key themes and quotes on a flip chart. All conversations were captured with a digital recorder. Thus, this paper is structured around three sources of data: a Qualtrics survey administered prior to the workshop (n=24), paper/audio recordings and transcriptions during the workshop, and a Qualtrics survey administered following the workshop (n=13). FEMA flood zones and land cover data were provided by First Street Flood Lab and USGS, respectively.

3. Results

3.1 Pre-workshop survey: Perceptions of flooding frequency

Figure 1 shows the pre-survey results for the question: "How frequent are the following types of floods?" in regards to rain-caused, ocean-caused and river-caused. To minimize confusion in terminology, rain flooding was described as "storm water, flash flooding, ponding or pluvial"; ocean flooding was described as "high tide flooding, king tide, storm surge, or coastal"; and river flooding was described as "flood plain flooding, over-topping banks, or fluvial". No one thought pluvial flooding was "not applicable" to their community/jurisdiction, and 37% felt it was "very frequent" or "constant". Thirteen percent of respondents believed that fluvial flooding did not apply to them. Of those that did, 20% felt this type of flooding was "very frequent" or "constant". Thirty-eight percent of respondents believed that tidal flooding did not apply to them. Of those that did, about the same percentage (19%) also placed this flooding into the same two highest categories. Interestingly, 79% of respondents believed that pluvial flooding had become more frequent over the past 10 years and no one thought it had become less frequent. This is compared to 56% (58%) of respondents who believed that fluvial (tidal) flooding had become more frequent over the past 10 years. Further, nine of the respondents believed that all three types of flooding were at least "somewhat frequent" in their community/jurisdiction and four of the nine believed all three flood types had become more frequent over the past 10-years. This speaks to the nontrivial threat of compound flooding in the study area.

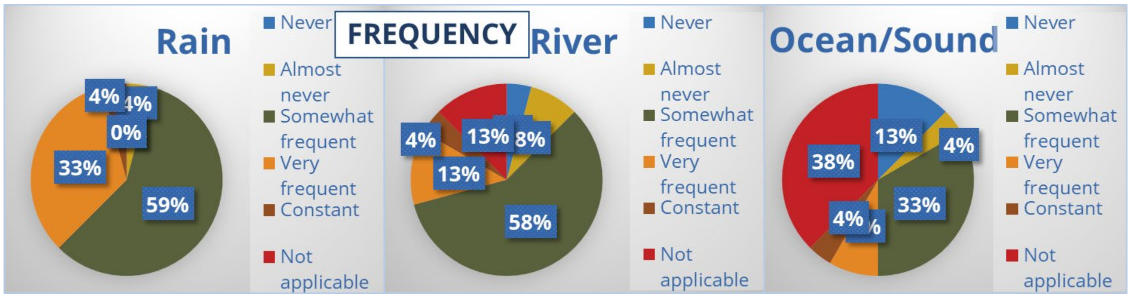


Fig 1. Responses to the question of the frequency of different types of flood.

3.2 Workshop data: Tyrrell County

As a case study, we isolated concerns from participants in Tyrrell County, NC. This low-lying rural county is constantly threatened by flooding. Nearly the entire county falls within the 100-year floodplain (Fig. 2a). About 7% of its citizens are employed in agriculture (ranking 7th out of 100 counties) and over 25% are living in poverty (ranking 5th). Over 28% of the land is classified as cultivated crops, but 55.5% are woody wetlands (Fig. 2b), which are mostly federally or state owned. As seen in Fig. 2b, there are distinct boundaries between these two land types. As expected, the wetlands are more flood prone than the agricultural land and this can lead to public-private tensions in flood management. As one participant put it: “53% of Tyrell County is owned by the state or federal government, who won’t let us touch it, who won’t go move a tree in it, and then wonder why we’re screaming about the fact our farmland is flooding…”

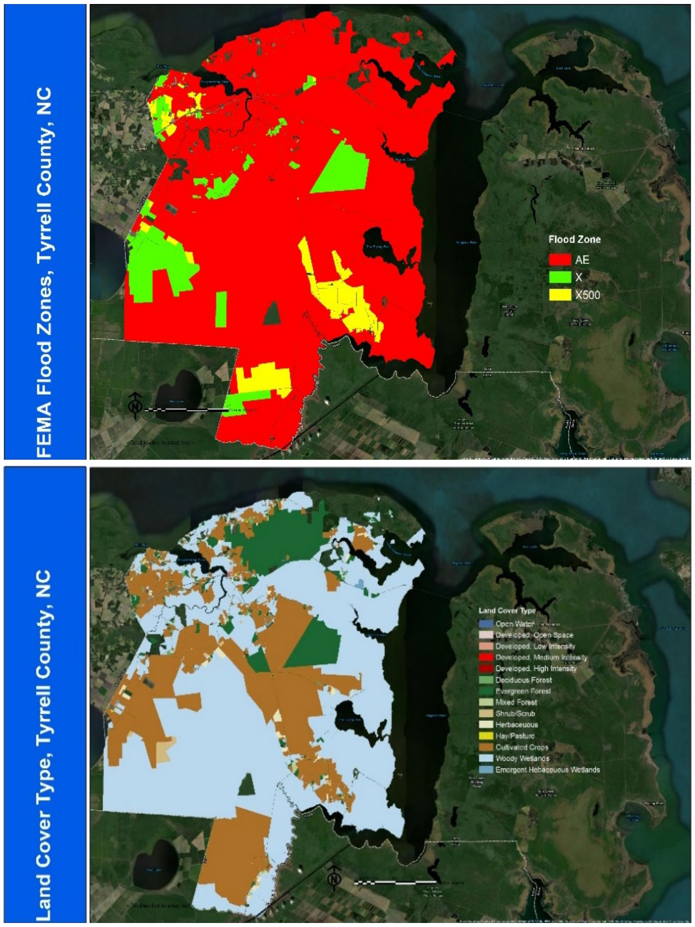


Fig. 2 (a) FEMA designated flood zones by property and (b) majority land use type by property in Tyrrell County, NC.

Another participant from Tyrrell described how the Soil and Water Conservation Districts have evolved in response to CCWE: “the Soil and Water Conservation Districts in every county have for years been more directed towards agriculture. It’s all about agriculture. They are slowly evolving what they see as their mission to a larger discussion, whether it be climate change or flooding or whatever. They need to be more in this discussion now, because used to they were all about agriculture. That was it. It’s a different world now, and they have accepted that. I’m just not so sure they have been viewed regionally for the expertise they bring to this discussion, because they have kind of transcended beyond agriculture. And particularly for Tyrrell, it’s rain-caused and it’s river-caused”.

3.3 Post-workshop survey: Assessment and COVID-19

Of the 13 participants who completed our post-workshop survey, 11 (85%) were moderately to extremely satisfied overall with the outcomes and all 13 would consider participating in the follow-up workshop (originally scheduled for 2021, but now slated for 2022). A couple of the participants wanted suggestions of flood prevention and mitigation measures, which will be a topic of discussion in the second workshop. There was also room for improvement in the facilitation and recording. As one participant observed: “some seemed knowledgeable and did a good job of capturing the concepts presented. Others seemed more unsure and the responses recorded on paper either missed a key point or didn’t capture the full breadth of the information shared”. Another participant thought moderators should have been more assertive in guiding the discussion or reeling in the focus to the topic at hand.

Given that the second survey was administered at the onset of the COVID-19 shut down, questions were included that asked how the pandemic might change the handling of flood hazard management either temporarily or permanently. Besides delays in implementing ordinances, working virtually and interacting with partners remotely, it was too early for most respondents to provide insights. However, one participant worried about the future: "What if this [shut down] had coincided with a flood event? What if people in isolation also had to be evacuated? What if the only staff in a county or municipality that understood the hazard mitigation grant process was also sick? Hurricane season is 8 weeks away. A significant flood could cause response and recovery issues in the flood plain that have never been thought of until now." His/her advice was for people to plan for the worst-case scenario and to emphasize a message of self-responsibility. "The combination of a pandemic and a flood event will overwhelm even the best system in a very short amount of time."

4. Conclusions

This paper presented some preliminary results from a NOAA-funded project: "Preparing for, Responding to, and Mitigating Coastal Compound Water Hazards for Resilient Rural Communities". Our sample of the hazard management and planning communities in eastern North Carolina thought that pluvial flooding was more pervasive and persistent than fluvial and tidal flooding. In the minds of many of the participants, this water hazard had also become more frequent over the past 10 years. Regarding the compound nature of floods (*i.e.* CCWE), the pre-workshop survey and tabletop discussions confirm that it is of growing concern.

During the workshop we asked whether there was cooperation across professional and jurisdictional boundaries to address CCWE risk. While most participants gave examples of functional partnerships, two themes on the importance of local knowledge and non-local governmental inflexibility did emerge. One case in point is Tyrrell County, NC where there is a disconnect in flood management between state and federally owned wetlands and adjacent privately-owned farm lands. Furthermore, the Soil and Water Conservation District is one source of local expertise that is not currently being exploited. Many more questions were explored in the workshop and we've identified ten key themes: flood causes, flood preparation, flood response, flood recovery, impacts, infrastructure, jurisdictional responsibility, networking and communication, planning and policies, and solutions. To view a mental map of the ten themes and keep up to date on project outcomes, we invite the reader to visit the project webpage: https://tinyurl.com/yyzzzz2t.

Finally, workshop participants were generally pleased with the event and wanted to continue the conversation. Plans are underway to hold a second workshop to satisfy objective 3 and "co-produce knowledge and tools with our study group for better preparation, response and mitigation plans". Given the concurrence of the pandemic with our project, we will also investigate this additional compounded hazard in the second workshop.

References

Cheng, S., E. Ganapati, and S. Ganapati, 2015: Measuring disaster recovery: Bouncing back or reaching the counterfactual state? *Disasters*, **39**, 427-446.

MDC, 2016: Building an infrastructure of opportunity. A REPORT FOR THE JOHN M. BELK ENDOWMENT. North Carolina's Economic Imperative. https://www.mdcinc.org/wp-content/uploads/2018/01/North-Carolinas-Economic-Imperative-Building-an-Infrastructure-of-Opportunity.pdf

Wahl, T., S. Jain, J. Bender, S. D. Meyers, and M. E. Luther, 2015: Increasing risk of compound flooding from storm surge and rainfall for major US cities. *Nat. Clim. Change*, **5**, 1093-1097.

Zscheischler, J., S. Westra, B. J. J. M. van den Hurk, S. I. Seneviratne, P. J. Ward, A. Pitman, A. AghaKouchak, D. N. Bresch, M. Leonard, T. Whal, and X. Zhang, 2018: Future climate risk from compound events. *Nat. Clim. Change*, **8**, 469-477.

NIST SP 1283
August 2022

**Proceedings of the ASME 2021
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE2021
August 17-20, 2021, Virtual, Online**

SEGMENTATION OF ADDITIVE MANUFACTURING DEFECTS USING U-NET

Vivian Wen Hui Wong[1], Max Ferguson[1], Kincho H. Law[1], Yung-Tsun Tina Lee[2], and Paul Witherell[2]

[1]Engineering Informatics Group, Civil and Environmental Engineering, Stanford University, Stanford, CA
[2]Systems Integration Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD

ABSTRACT

Additive manufacturing (AM) provides design flexibility and allows rapid fabrications of parts with complex geometries. The presence of internal defects, however, can lead to deficit performance of the fabricated part. X-ray Computed Tomography (XCT) is a non-destructive inspection technique often used for AM parts. Although defects within AM specimens can be identified and segmented by manually thresholding the XCT images, the process can be tedious and inefficient, and the segmentation results can be ambiguous. The variation in the shapes and appearances of defects also poses difficulty in accurately segmenting defects. This paper describes an automatic defect segmentation method using U-Net based deep convolutional neural network (CNN) architectures. Several models of U-Net variants are trained and validated on an AM XCT image dataset containing pores and cracks, achieving a best mean intersection over union (IOU) value of 0.993. Performance of various U-Net models is compared and analyzed. Specific to AM porosity segmentation with XCT images, several techniques in data augmentation and model development are introduced. This work demonstrates that, using XCT images, U-Net can be effectively applied for automatic segmentation of AM porosity with high accuracy. The method can potentially help improve quality control of AM parts in an industry setting.

Keywords: Smart Manufacturing, Defect Detection, Additive Manufacturing, Convolutional Neural Networks

1. INTRODUCTION

With years of development, additive manufacturing (AM), also known as 3D (three-dimensional) printing, has become an important technology in the manufacturing industry. The layer-by-layer process provides design flexibility and allows manufacturing of parts with complex geometries [1,2]. The fabrication process, however, comes with increased possibility of internal defects which are often difficult to detect. Layer-wise quality control is therefore very important for AM, as the internal defects could lead to undesirable properties in the fabricated part, resulting in deficit performance [3]. The ability to automatically identify defects of parts fabricated using AM is essential.

Current trends with non-destructive inspection (NDI) approaches often involve process monitoring through the installation of a large array of sensors and then analyzing and detecting failures using the collected sensor data [4,5,6]. These in-situ methods require the analyses of multiple signal types, and their correlations to final part quality are not yet well understood. Alternatively, ex-situ NDI techniques, such as X-ray computed tomography (XCT), are used to evaluate a completed build and offer a more reliable characterization of the AM part. XCT has emerged as perhaps the preferred technique for measuring properties of a completed AM build. It can be used to visualize internal structures and identify small pores and flaws in an AM part [7]. Obtaining useful images and segmentation labels from XCT scans, however, involves manual thresholding, making the process unscalable to a large number of samples.

Although many conventional methods to identify small defects remain difficult to implement in a manufacturing setting, the segmentation of defects in XCT images can be automated using computer vision and deep learning techniques. Here, the segmentation of AM defects refers to the ability to characterize and differentiate between porosity-indicative volumes and a fully dense part. Effective segmentation of defects enables more efficient identification, labeling, and sorting of such volumes. Defect segmentation can be framed as an image segmentation problem, which assigns each 2D pixel or 3D voxel of an image to a class. For defect segmentation, each pixel or voxel can be

Wong, Vivian; Ferguson, Max; Law, Kincho; Lee, Yung-Tsun; Witherell, Paul. "Segmentation of Addictive Manufacturing Defects Using U-Net." Presented at ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference IDETC/CIE 2021. August 17, 2021 - August 20, 2021.

classified as either the fully dense background or porosity using a deep learning model.

Convolutional neural networks (CNNs) have been commonly used for segmentation problems [8], and have been shown effective in many domains, including everyday objects [9], satellite imagery [10], and metal casting defects in manufacturing [11]. Most of these segmentation problems deal with 2D image data, but the biomedical domain, with its need to segment volumetric images such as computed tomography (CT) and magnetic resonance imaging (MRI) scans, poses the need for segmenting 3D images [12]. 3D CNNs have demonstrated potential in volumetric medical image segmentation [12,13,14]. Among existing 3D CNN methods, those with an encoder-decoder based architecture, also known as U-Net variants, have achieved excellent performance in several medical image segmentation tasks with relatively small number of training samples [15] and are increasingly popular in medical image segmentation applications, such as brain tumor segmentation [16] and 3D chest CT image segmentation for COVID-19 screening [17]. Images taken from metal AM parts, similar to their medical imaging counterparts, are volumetric and have comparable levels of contrast. Therefore, 3D CNNs that do well on medical image segmentation could possibly benefit AM image segmentation as well.

Despite the similarities between AM and medical imaging, AM presents many unique challenges. AM defects are pores and faults that are usually small (relative to the volumetric size of the part) and have highly irregular geometries. Furthermore, the sparsity of defects varies significantly between samples. In addition, because of the cost and manual effort needed to produce labeled AM datasets, very few AM datasets are publicly available. The lack of large public dataset poses a huge challenge to the adoption of machine-learning approaches, which require a large number of training data to converge on a reasonable model. Despite these challenges, the mainstream adoption of machine learning methods for defect detection of AM parts is essential to the developing of fast and reliable quality control procedures.

Motivated by the need for a defect segmentation method for quality control and inspired by the success of 3D CNNs in medical image segmentation, we applied 3D U-Net model with existing defect labels to automatically segment defects in XCT images of unknown AM samples [18]. In this paper, we focus on AM defects such as pores and cracks, or any internal, possibly defect-indicative, volumetric voids in the part. We show that 2D U-Net model, trained using 2D planar images, performs well and achieves high accuracy in terms of the mean intersection over union (IOU) measure. Our results demonstrate that both 2D U-Net and Residual 3D U-Net can reach high accuracy of 0.993 on the validation set. However, 2D U-Net may be better for some applications as it is easier to train and faster to evaluate. The contribution of this work is therefore not only to propose a method to automatically segment AM porosity with high accuracy, but also introduce techniques to augment the 3D U-Net models that can be used to directly perform porosity segmentation of a 3D volumetric part, which is particularly useful for AM parts that have complex geometries.

The rest of the paper is organized as follows: Section 2 provides an overview of related works. Section 3 describes the AM defect dataset that is used in this study. Section 4 describes background information on CNNs and the U-Net architecture, as well as the results in applying the U-Net models. Section 5 presents the approach taken to improve the performance of 3D U-Net models, including data augmentation and model development techniques. Finally, the paper is concluded with a brief summary and discussion in Section 6.

2. RELATED WORKS

With processing, 3D images can be sliced into 2D and vice versa, thereby allowing 2D CNNs to segment volumetric images [12]. The most commonly used CNN architectures for 2D segmentation problem are region-based and fully-convolutional-network-based (FCN-based) [8]. Region-based CNN (R-CNN), such as the Mask R-CNN, is an example of the former [8,9]. Since regions need to first be extracted, described, then classified, these methods are generally more computationally expensive [19]. On the other hand, FCN-based methods directly learn a mapping from input to output pixels, without proposing regions [20]. U-Net is a CNN model that extends the FCN architecture, achieving excellent performance, for example, in the segmentation of ventral nerve cord [21].

Despite the success of 2D CNN models, it has been suggested that since many medical images are 3D in nature, slicing them into 2D images prior to training loses information on the correlation between slices [22]. To that end, 3D FCN-based segmentation architectures, such as 3D U-Net [13] and V-Net [12], that train on volumetric medical images have been developed. While 3D CNN models can leverage information between slices, several disadvantages exist in comparison to 2D CNNs. 3D CNNs lack pre-trained models, leading to less stable training [22]. The patch-wise predictions in 3D are also more time-consuming to generate, compared to predictions in 2D. Furthermore, it has been pointed out that 2D U-Net may outperform 3D U-Net when the data is anisotropic [23]. To that end, in this study we compare the performance of 2D U-Net and 3D U-Net models on the same AM defect dataset.

A few related works have been done to automatically detect AM porosity using CNN. 2D detection and classification on porosity using slices of camera images have been reported [24]. CNN has also been used to analyze acoustic emissions during AM processes [25]. More recently, 2D U-Net has been deployed to a dataset similar to that used in this work and achieved a mean IOU of 0.92 [26]. This paper not only presents the novel CNN techniques for porosity segmentation, but also their applicability for achieving accurate results in the domain of AM XCT image segmentation, particularly, for 3D volumetric parts.

3. THE DATASET

The AM defect dataset used in this study was introduced by Kim et al. [27] and is publicly available [28]. The dataset consists of XCT images of four cobalt-chrome alloy, cylindrical AM specimens, created in a laboratory setting to investigate pore structures. Table 1 details the image size and porosity of each specimen. The metallic cylinders were produced using laser-based powder bed fusion (LPBF). Artificial pores and cracks were produced by changing AM scan speed and hatch spacing. By changing the process parameters, specimens were processed

TABLE 1: DETAILS OF THE AM DEFECT DATASETS

Specimen	Distance between 2D slices [pixel]	Image volume after 3D Reconstruction $(D \times W \times H)$ [pixel]	Porosity (%)
Sample 1	0.00245	$900 \times 980 \times 1010$	1.00
Sample 2	0.00277	$900 \times 988 \times 1013$	19.03
Sample 3	0.00243	$900 \times 984 \times 1010$	0.42
Sample 4	0.00252	$749 \times 984 \times 1010$	10.90

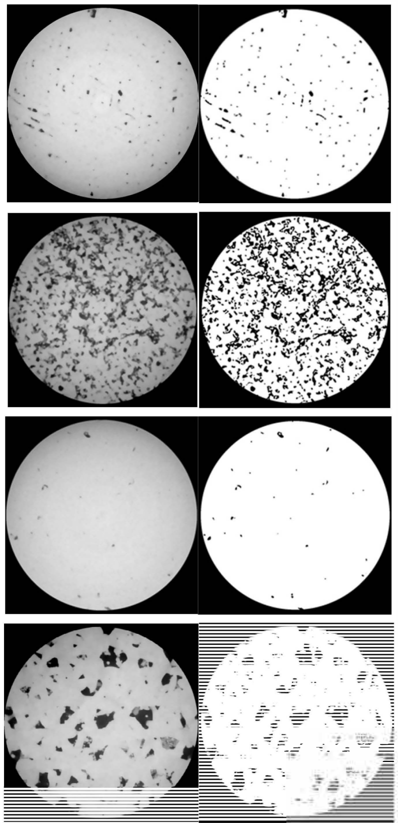


FIGURE 1: EXAMPLES OF IMAGES FROM AM DEFECT DATASETS: PROCESSED XCT IMAGES ARE ON THE LEFT AND SEGMENTATION MASKS ARE ON THE RIGHT

to have varying porosity. Then, XCT images of the specimens are taken. Each specimen's set of images consists of 8-bit grayscale images of 2D slices of XCT imagery. These images are 16-bit raw images obtained using XCT reconstruction processed by adding a $3 \times 3 \times 3$ median 3D filter and a non-local means filter [29,30]. To obtain ground truth labeling of the defects, Bernsen local thresholding [31] was used to process the 8-bit images. The local contrast threshold parameters of the thresholding process are computed by relating average noise value to local contrast threshold as explained by Kim et al. [27]. Figure 1 shows examples of images and corresponding labels.

The purpose of obtaining the XCT images and thresholding for the labeling of defects is to use them as inputs and ground truth reference for CNN models. CNN models use the images as inputs and produce predicted segmentation masks, classifying defect and background pixels or voxels, and compare the prediction results with the ground truth to obtain a loss function, which is then minimized through an iterative training process.

In order to evaluate the applicability of 3D CNN model for volumetric images, the 2D XCT images are concatenated into a 3D volumetric image, restoring the original cylindrical form of an AM sample, as shown in Figure 2. The AM defect dataset has the following characteristics to be noted:

- The standard deviations of the pores on the z-axis of some samples are not the same as those of the x and y axes, meaning that the shape and distribution of pores may be **anisotropic** [27].
- The 3D structure of the defects gives limited information about the location according to the ground truth labels, as the labels are generated by thresholding 2D images.
- As shown in Figure 1, the geometries of the defects can look highly **irregular**.
- Percentage of porosity indicates that there is an **imbalance** in the number of porosity and background voxels.

Altogether four specimens with images are available for the study. Three specimens (samples 2, 3 and 4) are used for training and one specimen (sample 1) is used for validation of the trained models. Since the samples range vastly in porosities, the validation sample is selected because its percentage of porosity is neither the minimum nor the maximum.

4. CONVOLUTIONAL NEURAL NETWORKS (CNNS) AND U-NET

Developments in CNNs in the past decade have significantly improved the ability to perform image classification, detection and segmentation in many domains. This section first gives a brief overview of deep CNNs. We then introduce the U-Net architecture, which is the architecture that inspires many of recent domain-specific works on image segmentation.

4.1 Convolutional Neural Networks

A CNN is a type of deep neural network that consists of several layers, where each layer uses mathematical operations, such as convolution, to convert the input to a feature map. CNNs have been commonly employed and operated on 2D images and have recently been extended to the study of 3D images. The idea of training a 2D or a 3D CNN model is identical, but with the following distinctions:

1) The convolving kernels in 3D CNNs are 3D with width, height and depth $(W \times H \times D)$, whereas the kernels in 2D CNNs are 2D with width and height $(W \times H)$ only.
2) When convolving, a 3D kernel moves in 3 directions, along all 3 axes of the input image and its feature maps. A 2D kernel moves in 2 directions, along the axes corresponding to W and H dimensions.

Figure 3 shows an example a 3D CNN architecture with multiple types of layers. A layer l of a neural network can be written as a function parameterized by parameters $\boldsymbol{\theta}^{(l)}$:

$$h^{(l)} = f^{(l)}\left(h^{(l-1)}; \boldsymbol{\theta}^{(l)}\right) \tag{1}$$

where $h^{(l)}$ is the output feature map of layer l and $h^{(0)}$ is an

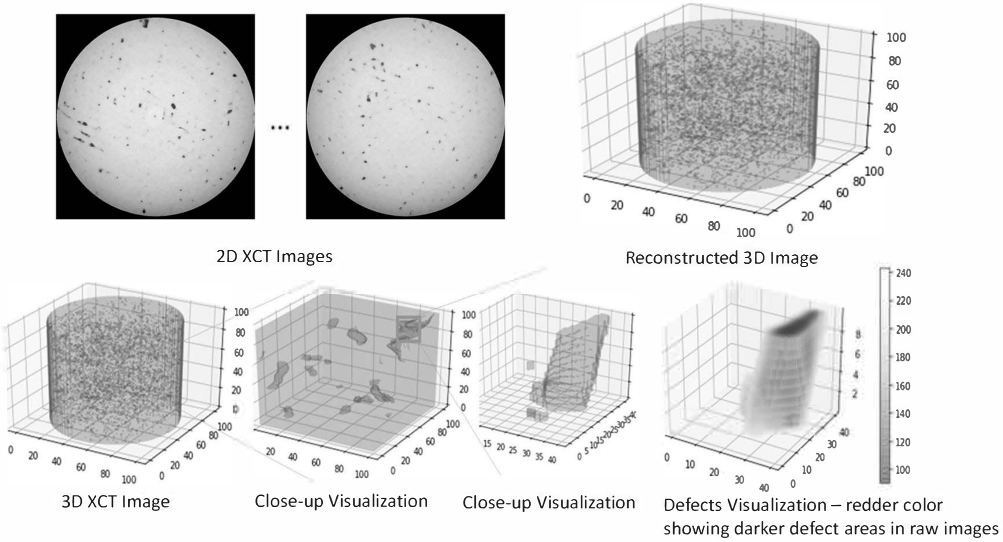


FIGURE 2: RECONSTRUCTION OF A 3D AM IMAGE FROM 2D XCT IMAGES.

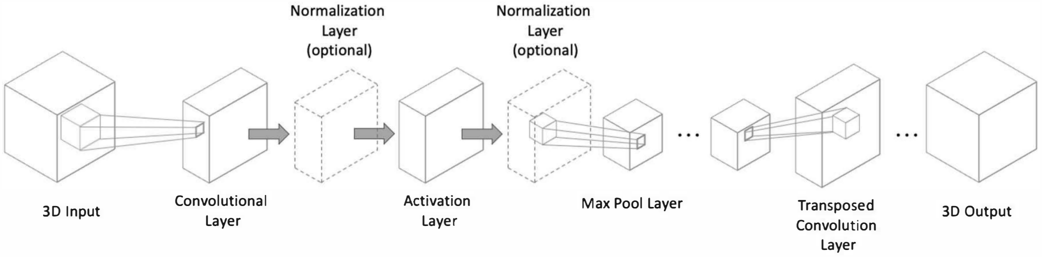

Normalization Layer (optional)
Normalization Layer (optional)
3D Input
Convolutional Layer
Activation Layer
Max Pool Layer
...
Transposed Convolution Layer
...
3D Output

FIGURE 3: EXAMPLE OF A 3D CNN BUILT BY COMBINING VARIOUS TYPES OF LAYERS

image tensor. A layer can be a convolutional layer, which uses a parameterized kernel to convolve with the layer's input. In a convolution layer, the dot product of the kernel and the input at each spatial location is taken. The stacked layers approximate a complex function of the image input and the output $h^{(L)}$ at the final layer L representing the model's prediction. With the stacking of multiple layers, a parameterized function mapping the input image to the prediction can be created. To introduce nonlinearity in the function approximator, a nonlinear activation function is commonly applied to the output of a layer:

$$h^{(l)} = \sigma\left(f^{(l)}\left(h^{(l-1)}; \boldsymbol{\theta}^{(l)}\right)\right) \qquad (2)$$

where σ is a nonlinear function. Common choices for σ are the sigmoid function and the rectified linear unit (ReLU) function.

Another type of layer is a normalization layer, such as batch normalization (BN) or group normalization (GN) layer. A normalization layer normalizes its input in order to improve the speed and stability of training neural networks [32]. BN performs the normalization along the batch and spatial locations. On the other hand, GN, which is more robust than BN, divides the input's channels into groups and performs normalization for each group. GN alleviates the limitation of BN that smaller batch size leads to larger errors [33]. Instance normalization (IN) is another technique that normalizes across spatial locations [34].

Max pool and transposed convolution layers can also be used in CNN models to respectively downsample or upsample the input tensor spatially. A max pool kernel draws the maximum at each of the input's spatial locations. A transposed convolution multiplies the input at each spatial location with the kernel and adds the result to the layer's output at the same location.

A deep neural network is trained by minimizing a loss function. The loss function measures the amount to which the prediction differs from the ground truth. During training, the model's parameters, $\boldsymbol{\theta} = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^L\}$, are updated using the gradient of the loss function, which, thereby, must be differentiable. This method of calculating gradients with respect to the parameters is generally known as backpropagation [35].

4.2 2D AND 3D U-NET MODELS

Due to its excellent performance, U-Net is a popular CNN architecture not only in the medical [16,17,21] and but also in non-medical domains such as satellite imagery [36]. The design of U-Net is characterized by two properties: The U-shaped structure formed by an encoder and a decoder network, as well as the skip connections that connect the corresponding

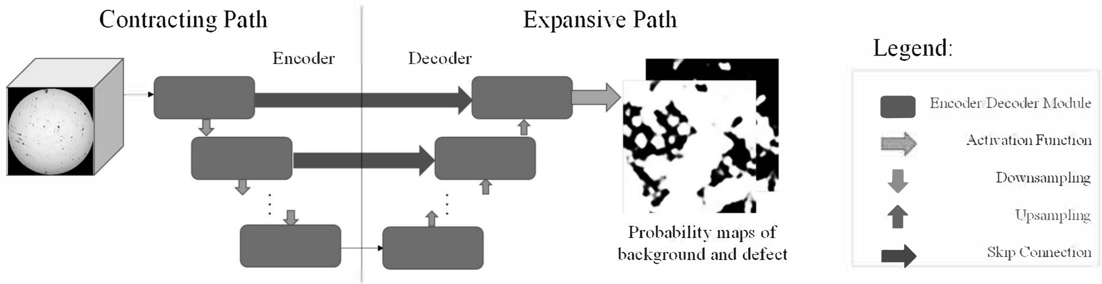


FIGURE 4: SCHEMATIC OF A GENERAL U-NET ARCHITECTURE

TABLE 2: DETAILS OF U-NET BASE CONFIGURATION

Model	Encoder/Decoder	Downsampling	Upsampling
2D U-Net	$(3 \times 3$ convolution +BN + ReLU$) \times 2$	2×2 max pool	Bilinear with a scale of 2
Vanilla 3D U-Net	$(3 \times 3 \times 3$ convolution +BN + ReLU$) \times 2$	$2 \times 2 \times 2$ max pool	$2 \times 2 \times 2$ transposed convolution
3D U-Net with GN	$(3 \times 3 \times 3$ convolution +ReLU+ GN$) \times 2$		
Residual 3D U-Net	$(3 \times 3 \times 3$ convolution +GN+ ReLU$) \times 3$		

encoders and decoders. Figure 4 shows the overall structure of a U-Net architecture. Implementation in this work closely follows this general structure, but with changes in design details of the encoders, decoders, upsampling and downsampling.

The U-Net architecture consists of two main parts. The contracting path on the left consists of encoder modules and downsampling operations that increase the number of feature maps produced as the number of layers increases. The expansive path on the right consists of decoder modules and upsampling operations that decrease the number of feature maps as the number of layers increases. Although other variations (such as normalization) exist, the encoder and decoder modules are normally convolutions with activation functions. The encoder and decoder modules that have the same resolution are connected with a skip connection, combining their outputs to produce the input for the next decoder module.

The U-Net architecture's modular design allows for flexibility in altering its modules. The classical 2D U-Net's encoder and decoder modules are double convolutions using 2D kernels and a ReLU activation [21]. On the other hand, 3D U-Net, described in [13], deploys 3D convolutions and adds batch normalization to its encoder and decoder modules. Furthermore, the ResUNet-a presented in [37] discovered that residual connections could improve the performance of U-Net and can help reduce the vanishing gradient problem [38]. Therefore, in ResUNet-a, residual connections were added to the encoder and decoder modules of the U-Net architecture. Combining residual connections and 3D U-Net has been shown to perform well in several medical imagery segmentation tasks [14,38].

4.3 IMPLEMENTATION

To assess the performance of 2D and 3D U-Net on AM porosity segmentation, we train and evaluate several U-Net configurations, namely, the 2D U-Net, vanilla 3D U-Net, 3D U-Net with GN (instead of BN in the vanilla configuration) and 3D U-Net with residual connections, as summarized in Table 2.

The 2D U-Net follows the same implementation as described in [21]. The model's encoder and decoder modules are each a double convolution: twice stacking a 2D convolutional layer, followed by a ReLU activation. 3×3 kernels are employed for the convolutional layers, and 2×2 max pooling is used for downsampling. Some minor modifications from the standard implementation are made: First, bilinear upsamplings are used instead of transposed convolution to save memory. Furthermore, the 2D convolutions are zero padded with a one-pixel border to preserve the features of the edges. Lastly, a BN layer is added after each 2D convolutional layer and before the ReLU to improve stability. Inputs to the 2D U-Net are the raw 2D images and masks as given in the dataset. The training minimizes cross- entropy loss with a RMSprop optimizer [39], and a learning rate of 0.0001 is used. The 2D U-Net implementation is adapted from a publicly available PyTorch implementation [40].

The vanilla 3D U-Net, on the other hand, follows the implementation by Cicek et al. [13]. The model follows a similar architecture as the 2D U-Net, but with 3D convolutions. The double convolutional layers consist of a 3D convolutional layer, followed by a BN layer and a ReLU nonlinearity layer, all stacked twice to form the double convolution. The second 3D U-Net model is a variant that uses GN as the normalization layer and places the GN layer after the ReLU activation layer. The third model, Residual Symmetric 3D U-Net, follows the implementation by Lee et al. [14], which introduces residual skip connections in the modules and modifies the upsampling and downsampling techniques. Inputs to the 3D U-Net models are 3D images, constructed by stacking the 2D slices as described in Section 3. Due to memory constraint, each training sample is a 128×128×128 patch randomly sampled from the 3D image. Stride sizes are 32×32×32 to overlap the patches and ensure that information is not lost. The input patches are normalized, randomly flipped and rotated prior to training. Network outputs and targets are compared using the cross-entropy loss. Each model is trained with an initial learning rate of 0.0002 that decays at a rate of a half at the 600th, 1000th, and 1400th iterations. The networks are trained via the Adam optimizer [41]. A weight decay factor of 0.0001 is used. The batch size and the group size are set to one for BN and GN layers, respectively. All modifications to the models are conducted using a publicly available implementation of the 3D U-Net architecture [42]. All models are trained on NVIDIA Tesla T4 GPUs with 100 GB RAM and 4 Intel virtual CPU on Google Cloud Platform.

4.4 EXPERIMENTAL RESULTS

The prediction accuracy of each of the above-mentioned models is evaluated using the mean IOU metric, comparing the accuracy of a predicted segmentation with the ground truth or labeled mask. Table 3 shows the mean IOU and the training time to achieve the accuracy for the AM datasets. The 2D U-Net outperforms the 3D U-Net models, which could be due to the fact that our dataset is anisotropic and thus, as suggested in [20], favors the performance of 2D U-Net. Among the 3D models, the Residual 3D U-Net model requires the longest training time but performs slightly better than the other 3D U-Net models.

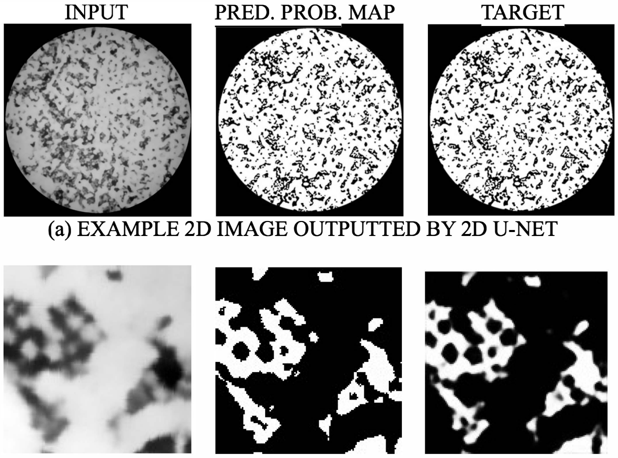
We observe several limitations posed by the 3D models. Figure 5 shows an example patch sampled by the Residual 3D U-Net from the validation data. It can be seen that due to the large size of the images, the $128 \times 128 \times 128$ patches only capture the shape of defects partially. This could limit the model's ability to predict based on the defect's relative spatial location. Sharp edges of the irregular defects are often misclassified, which could be an indication that the additional axis of information that 3D CNNs leverage does not compensate for the loss of global information in the $W - H$ plane. We also provide a 2D image segmented by the 2D U-Net in Figure 5, showing that the 2D U-Net's predictions are mostly accurate.

Some drawbacks can also be observed on both the 2D and 3D models. The challenges observed from the predictions using the AM defect dataset are as follows:

1) **Variation in sizes:** AM defects can range from hardly visible, very small voids to large voids, as shown in Figure 1. Small defects are challenging for segmentation using CNN because there are inherently less voxels of these smaller defects for training, and they are difficult to distinguish from background noises.
2) **Lack of training voxels:** Following the previous point, there are much fewer defect voxels than the background voxels, which can cause complications. The background voxels located outside the rim of the cylinder are considered trivial. However, because they are naturally dark, they are always thresholded as defects. As shown in Table 1, the specimen with the highest porosity has less than 20% porosity, meaning that there is a significant class imbalance in the training examples.
3) **Highly irregular geometry:** It can be viewed in Figure 1 that the shapes of defects are highly irregular, often consisting of sharp edges and light color rims. This poses difficulties for a CNN model to infer the correct geometry from surrounding voxels, and the boundaries of such irregular shapes are difficult to identify.
4) **High resolution:** The resolution size of the input images in the referenced AM dataset is very large. Table 1 shows the number of voxels and the array shape of each specimen image. This not only leads to a high memory consumption during preprocessing and training, but also discards the possibility to conduct downsampling prior to training, as downsampling would lose valuable information on the already rare small defects.

TABLE 3: VALIDATION MEAN IOU AND AVERAGE TRAINING TIME PER EPOCH ON THE BASE U-NET MODELS

Model	Training Time (hours)	Validation mean IOU
2D U-Net	**0.70**	**0.993**
Vanilla 3D U-Net	6.58	0.863
3D U-Net with GN	14.00	0.881
Residual 3D U-Net	19.97	0.884



(a) EXAMPLE 2D IMAGE OUTPUTTED BY 2D U-NET

(b) EXAMPLE PATCH OF A DEFECT OUPUTTED BY RESIDUE 3D U-NET

FIGURE 5: EXAMPLES OF SEGMENTATION RESULTS OUTPUTTED BY 2D U-NET AND RESIDUE 3D U-NET

In the next section, we propose a number of enhancements on the dataset to improve the performance of the U-Net models.

5. DATA AUGMENTATION AND MODEL DEVELOPMENTS

Although the results in Table 3 show that 2D U-Net outperforms 3D U-Net on the dataset used in this study. A 3D U-Net model could be useful to directly make predictions on AM volumetric parts with complex geometries. This section describes an approach that can enhance the performance of 3D U-Net models on segmenting AM defects based on the nnU-Net, which is a framework that performs preprocessing, U-Net configuration, training and post-processing for image segmentation [43]. Figure 6 shows how choices of data enhancement techniques can help address the four identified challenges of AM defect segmentation. We also perform the same techniques on the 2D U-Net to observe their effects. The five techniques as shown in Figure 6 are as follows:

- **Nonzero Cropping:** The images are cropped into regions where voxel values are nonzero. Each AM specimen image consists of a large number of voxels. Such large arrays of voxels induce a heavy computational cost. Cropping the images can reduce the size of arrays while keeping the data containing valuable information for training. However, since convolution is done on rectangular images, some voxels outside of the cylinders are preserved after cropping.

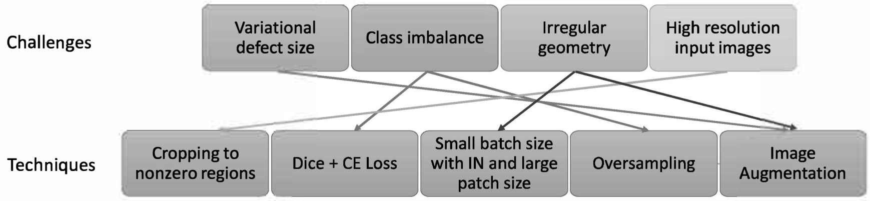


FIGURE 6: TECHNIQUES USED TO ADDRESS SEGMENTATION CHALLENGES POSED BY AM DEFECTS

- **Dice + Cross-Entropy (CE) Loss:** A loss function that sums the Dice loss and cross-entropy (CE) loss is used [44]. Dice loss is a commonly used loss function for segmentation. Here, the objective function is set as the Dice score, an evaluation metric for accuracy that considers class imbalance. However, since training is done in patches, we cannot calculate the Dice score of the entire image based on any single patch. An estimated Dice score, derived from combined patches, could be an inaccurate estimate of the true Dice score and lead to unstable training. On the other hand, cross-entropy (CE) loss, another commonly used loss function, measures the degree at which the prediction differs from the true label. It is found empirically that combining CE loss and Dice loss improves segmentation quality [45]. Therefore, to address both training stability and the imbalanced number of defects and background voxels, the loss function is selected to be the sum of the Dice loss and the cross-entropy (CE) loss functions as follows:

$$L_{dc} = 1 - \frac{\sum_{i \in I} h_i y_i}{\sum_{i \in I} h_i + \sum_{i \in I} y_i} \tag{3}$$

$$L_{ce} = -(y \log(h) + (1 - y) \log(1 - h)) \tag{4}$$

$$L_{total} = L_{dc} + L_{ce} \tag{5}$$

  where L_{dc} is the Dice loss averaged over all batches, L_{ce} is the cross-entropy loss, L_{total} is the total loss, h is the model prediction and y is the ground truth.
- **Batch and patch size:** Classification of voxels at boundaries of irregular defects is a difficult task for the network. Typically, a larger training patch size means that more contextual information from surrounding voxels are incorporated when computing the weights. Capturing the full shape of a defect could also lead to less confusion over the boundaries. A larger patch size is therefore desirable in training but results in a reduction of the batch size. For this reason, the batch size is selected to be small. While batch normalization [32] is often used in CNN training to improve robustness and convergence, but, because of the smaller batch size, instance normalization [34] is used. Lastly, the size of kernels is calculated by limiting the total size of feature maps to the GPU memory budget.
- **Oversampling:** Since we have less voxels of defects than voxels of background, the imbalance in training data leads to the lack of training data, thereby impacting the accuracy of segmentation. Oversampling resolves the issue by sampling examples containing the rarer class, in this case the defects, more often than the more dominant class. When sampling patches for training, we ensure that the rarer class is sampled more frequently: Patches are sampled such that at least one of the patches or one third of the patches in a batch, whichever is greater, is guaranteed to contain a randomly selected defect voxel, with the rest of the batch being randomly sampled.
- **Image Augmentation:** Multiple data augmentation techniques are used during training. Images used as inputs to the network are normalized to ensure that each voxel has a similar distribution. This adds robustness and improves convergence during training. In our approach, each image is normalized by subtracting the mean and dividing by the standard deviation of voxels in that image. Images are randomly rotated and scaled. Furthermore, we add additional noise into the dataset to improve robustness. With a certain probability controlled by a random number generator, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, Gamma augmentation and mirroring are applied. These augmentation techniques help the network to generalize defects with various aspect ratios, colors, and shapes. Details and specifics of these augmentation techniques have been described by Isensee et al. on nnU-Net [43].

In addition to the data enhancements made above, several minor modifications to the original U-Net architecture are made. The ReLU activation functions are replaced with Leaky ReLU, and downsampling is implemented as strided convolution. Deep supervision is used in training, which adds an additional term for loss in some larger feature maps of the decoder. These modifications are useful design choices to facilitate training [43].

In the last step of the approach, we train a CNN model that utilizes the five data enhancement techniques and the architectural design choices mentioned previously. To train a model, we sample minibatches and train iteratively to optimize the layer parameters over the Dice + CE loss function.

5.1 IMPLEMENTATION

The models implemented with data enhancement techniques are shown in Table 4. All models are trained end-to-end and without pretraining, with weights initialized using the initialization procedure described by He et al. [46]. Stochastic gradient descent with Nesterov momentum [47] is used to optimize the learning. The initial learning rate is selected at 0.01, and decays throughout training at a rate of 9×10^{-6} per epoch. Each epoch is defined as 250 training iterations on the minibatches. The total number of epochs is determined based on the convergence of losses. The training loss is calculated by summing cross-entropy loss and batch Dice loss. Since trade-off

TABLE 4: DETAILS OF ENHANCED U-NET CONFIGURATION

Model	Encoder/Decoder	Downsampling	Upsampling
2D U-Net (Patched)	(3 × 3, stride 2 convolution + IN + Leaky ReLU) × 2	Done through strided convolution	2 × 2 transposed convolution
3D U-Net	(3 × 3 × 3, stride 2 convolution +IN + Leaky ReLU) × 2		2 × 2 × 2 transposed convolution
Residual 3D U-Net	(3 × 3 × 3, stride 2 convolution +IN + Leaky ReLU) × 2		

exists between runtime and loss reduction, training is terminated at 56 epochs, when all models have reached a plateau in losses.

Given the goal of a small batch size, and constrained by the GPU's capacity, the 3D U-Net and the Residual 3D U-Net use a batch size of 2, with each patch size being 128 × 128 × 128. The 2D U-Net uses a batch size of 3, with the patch size of 1024 × 1024. Input images are augmented using the previously mentioned image augmentation techniques conducted on the fly during the training process. The inference procedure is patch-based and uses the same patch size used during training.

5.2 Experimental Results

The performances of the three enhanced U-Net models are shown in Table 5 where the mean IOU evaluation scores and the amount of time taken for training are reported. As shown in the table, the Residual 3D U-Net model, with a mean IOU of 0.993, achieves the highest accuracy, and is comparable to the non-patched 2D U-Net as shown in Table 3. The training of the 3D models requires, as expected, much more time than the 2D counterpart, which, with patch-based sampling, also takes longer time than the original 2D U-Net model.

Figure 7 shows a slice of the segmentation mask outputted by the Residual 3D U-Net model. It can be observed that most defects have been segmented by the model. The prediction resembles well with the labeled mask and is able to segment the complex geometries of most identified defects. However, it should be noticed that defects with very light colors in the input have more ambiguous labels, and therefore those voxels may not necessarily be classified correctly by the model.

6. SUMMARY AND DISCUSSION

This paper presents the use of U-Net models for automatic detection of AM defects using XCT images. Using the dataset available in this study, the 2D U-Net achieves accurate segmentation with the shortest training time. Although the 2D U-Net seems to be the best fit for this AM defect dataset, one must note that the dataset contains several characteristics as described in Section 3, which could lead to the 2D U-Net outperforming the 3D U-Net models. 3D U-Net models may become more effective with other datasets and scenarios, for example, when the geometry of the AM fabricated part is complex, or when the CT images are much noisier. In practice, AM parts have more complex geometry than the cylindrical specimens employed in this study. 3D model would allow better differentiation between intended and un-intended porosity, for example, for parts that have internal features such as holes and channels inside.

With minor modifications in network architectures, the mean IOU increased substantially for the 3D U-Net models. We attribute the improved accuracy to the various data enhancement techniques used, including additional preprocessing, oversampling, image augmentation, as well as the change in the design of the loss function. These techniques are purposely tailored towards improving prediction given the domain-specific challenges. We argue that these techniques take on a more significant role than minor changes in network architecture design when deploying 3D U-Net models on the same dataset and suggest that these techniques should be considered in future related works to improve model performance. Furthermore, for situations that deem necessary, attention modules can be introduced to potentially further improve model accuracy [48].

TABLE 5: VALIDATION MEAN IOU AND AVERAGE TRAINING TIME PER EPOCH ON ENHANCED U-NET MODELS

Model	Training Time (hours)	Validation mean IOU
2D U-Net (Patched)	4.55	0.988
3D U-Net	21.06	0.992
Residual 3D U-Net	20.61	**0.993**

In summary, while conventional manual or thresholding methods for AM defect segmentation remain tedious and unscalable, this paper has presented a method for automatic volumetric segmentation of AM specimens -- a challenging task given: the complex geometries of the specimens, the poor contrast and lighting resulting from measuring metal specimens, and the imbalance of defect and background classes associated with the images. A high predictive accuracy with a mean IOU of 0.993 is achieved by the 2D U-Net model and the Residual 3D U-Net model with data enhancements. The high accuracy of the method demonstrates the potential of deep learning models to be applied to aid the quality control of AM parts in practice.

ACKNOWLEDGEMENTS

This research is partially supported by the Measurement Science for Additive Manufacturing 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 3D 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

[1] Gibson I., Rosen, D., and Stucker, B., 2010, *Additive Manufacturing Technologies*, Springer, New York.

[2] Ngo, T. D., Kashani, A., Imbalzano, G., Nguyen, K. T. Q., and Hui, D., 2018, "Additive manufacturing 3D printing: a

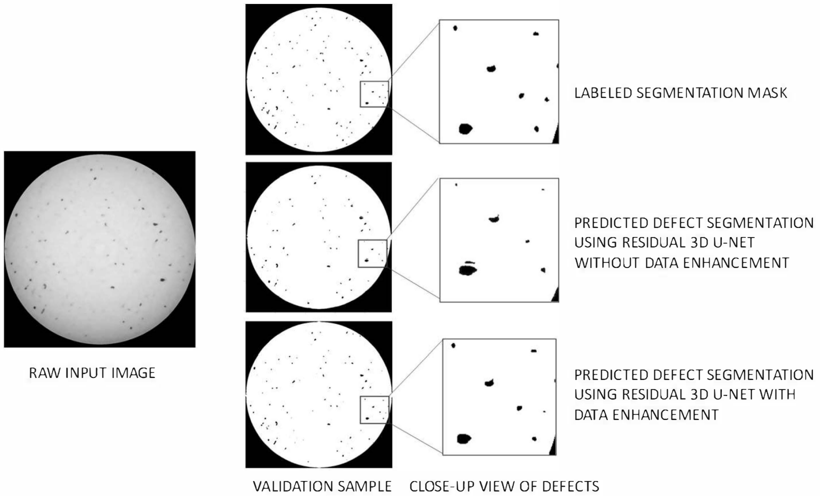

LABELED SEGMENTATION MASK
PREDICTED DEFECT SEGMENTATION USING RESIDUAL 3D U-NET WITHOUT DATA ENHANCEMENT
PREDICTED DEFECT SEGMENTATION USING RESIDUAL 3D U-NET WITH DATA ENHANCEMENT
RAW INPUT IMAGE
VALIDATION SAMPLE
CLOSE-UP VIEW OF DEFECTS

FIGURE 7: 2D SLICE OF A SAMPLE SEGMENTED BY RESIDUAL 3D U-NET WITH AND WITHOUT DATA ENHANCEMENT

review of materials, methods, applications and challenges," *Composites Part B: Engineering*, **143**, pp.172–196.

[3] Reese, R., Bheda, H., and Mondesir, W., 2016, "Method to monitor additive manufacturing process for detection and in-situ correction of defects," *Pub. No.: US 2016/0271610 A1 Patent Application Publication*.

[4] Wu, H., Wang, Y., and Yu, Z., 2016, "In situ monitoring of FDM machine condition via acoustic emission," *International Journal of Advanced Manufacturing Technology*, **84**(5-8), pp. 1483-1495.

[5] Faes, M., Abbeloos, W., Vogeler, F., Valkenaers, H., Coppens, K., Goedemé, T., and Ferraris, E., 2016, "Process monitoring of extrusion based 3D printing via laser scanning," *arXiv preprint arXiv:1612.02219*.

[6] Rao, P.K., Liu, J.P., Roberson, D., and Kong, Z.J., 2015, "Sensor-based online process fault detection in additive manufacturing," *ASME 2015 International Manufacturing Science and Engineering Conference*, ASME Digital Collection, V002T04A010-V002T04A010.

[7] 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): pp. 115–126.

[8] 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): pp. 87–93.

[9] He, K., Gkioxari, G., Dollár, P., and Girshick, R., 2017, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Venice, Italy, pp. 2980–2988.

[10] Pesaresi, M., and Benediktsson, J. A., 2001, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, **39**(2), pp. 309-320.

[11] 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), pp. 137-164.

[12] 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,* pp. 565-571.

[13] Cicek,O., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O., 2016, "3D U-Net: learning dense volumetric segmentation from sparse annotation," *International conference on medical image computing and computer-assisted intervention (MICCAI)*, pp. 424–432.

[14] 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,*.

[15] Singh, S.P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., and Gulyás, B., 2020, "3D Deep Learning on Medical Images: A Review," Sensors, **20**(18), pp.5097.

[16] Henry, T., Carre, A., Lerousseau, M., Estienne, T., Robert, C., Paragios, N., and Deutsch, E., 2020, "Top 10 BraTS 2020 challenge solution: Brain tumor segmentation with self-ensembled, deeply-supervised 3D-Unet like neural networks," *arXiv preprint arXiv:2011.01045*.

[17] Wang, J., Bao, Y., Wen, Y., Lu, H., Luo, H., Xiang, Y., Li, X., Liu, C., and Qian, D., 2020, "Prior-attention residual learning for more discriminative COVID-19 screening in CT images," *IEEE Transactions on Medical Imaging*, **39**(8), pp.2572-2583.

[18] Wong, V.W.H., Ferguson, M., Law, K.H., Lee, Y.-T.T. and Witherell, P. 2020, "Automatic volumetric segmentation of additive manufacturing defects with 3D U-Net," *AAAI 2020 Spring Symposia*, Stanford, CA, USA, Mar 23-25, 2020. *arXiv preprint arXiv:2101.08993***.**

[19] Caesar, H., Uijlings, J., and Ferrari, V., 2016, "Region-based semantic segmentation with end-to-end training," *arXiv preprint arXiv 1607.07671.*

[20] Long, J., Shelhamer, E., and Darrell, T., 2015, "Fully convolutional networks for semantic segmentation," *IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

[21] Ronneberger, O., Fischer, P., and Brox, T., 2015, "U-Net: convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241.

[22] 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.*

[23] Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S, and Maier-Hein, K.H., 2017, "Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features," *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 120-129.

[24] Zhang, B., Liu, S., and Shin, Y. C., 2019, "In-process monitoring of porosity during laser additive manufacturing process," *Additive Manufacturing*, **28**, pp. 497–505.

[25] 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**, pp. 598–604.

[26] Mutiargo, B., Pavlovic, M., Malcolm, A.A., Goh, B., Krishnan, M., Shota, T., Shaista, H., Jhinaoui, A., and Putro, M.I.S., 2019, "Evaluation of X-Ray computed tomography (CT) images of additively manufactured components using deep learning," *3rd Singapore International Non-destructive Testing Conference and Exhibition (SINCE2019).*

[27] 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**, pp. 23-38.

[28] 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. Available at https://doi.org/10.18434/M32162. (Accessed 11/2019).

[29] Buades, A., Coll, B., and Morel, J-M., 2011, "Non-local means denoising," *Image Processing On Line*, **1**, pp. 208-212.

[30] 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.

[31] Bernsen, J., 1986, "Dynamic thresholding of gray-level images," *8th International. Conference on Pattern Recognition*, pp. 1251–1255.

[32] Ioffe, S., and Szegedy, C., 2015, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167.*

[33] Wu, Y., and He, K., 2018, "Group normalization," *arXiv preprint arXiv:1803.08494*

[34] Ulyanov, D., Vedaldi, A., and Lempitsky, V., 2016, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022.*

[35] Werbos, P. J., 1990, "Backpropagation through time: what it does and how to do it," **78**, pp. 1550–1560.

[36] Rakhlin, A., Davydow, A., and Nikolenko, S.I., 2018, "Land cover classification from satellite imagery with U-Net and Lovasz-softmax loss," *CVPR Workshops*, pp. 262-266.

[37] Diakogiannis, F.I., Waldner, F., Caccetta, P. and Wu, C., 2020, "ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, **162**, pp. 94-114.

[38] He, K., Zhang, X., Ren, S., and Sun, J., 2016, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770–778.

[39] Hinton, G., 2012, *Neural Networks for Machine Learning - Lecture 6a - Overview of Mini-Batch Gradient Descent.* Available at https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf. (Accessed 7/2020)

[40] *UNet: Semantic Segmentation with PyTorch*, Available at https://github.com/milesial/Pytorch-UNet (Accessed 11/2020).

[41] Kingma, D., and Ba, J., 2014, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980.*

[42] Wolny, A., 2019, "Wolny/Pytorch-3DUnet: PyTorch implementation of 3D U-Net," *Zenodo.* http://doi.org/10.5281/ zenodo.2671581

[43] Isensee, F., Jäger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H., 2020, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128.*

[44] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C., 2016, "The importance of skip connections in biomedical image segmentation," *Deep Learning and Data Labeling for Medical Applications*, pp. 179-187.

[45] Khened, M., Kollerathu, V.A., and Krishnamurthi, G., 2019, "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Medical image analysis*, *51*, pp.21-45.

[46] He, K., Zhang, X., Ren, S., and Sun, J., 2015, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026-1034.

[47] Yurii, N., 2013, *Introductory Lectures on Convex Optimization: A Basic Course*, Springer Science & Business Media.

[48] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., and Glocker, B., 2018, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999.*

Scaling the Phish: Advancing the NIST Phish Scale

Fern Barrientos, Jody Jacobs(✉), and Shaneé Dawkins

National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
{fernando.barrientos,jody.jacobs,shanee.dawkins}@nist.gov

Abstract. Organizations use phishing training exercises to help employees defend against the phishing threats that get through automatic email filters, reducing potential compromise of information security and privacy for both the individual and their organization. These exercises use fake and realistic phishing emails to test employees' ability to detect the phish, resulting in click rates which the organization can then use to address and inform their cybersecurity training programs. However, click rates alone are unable to provide a holistic picture of why employees do or do not fall for phish emails. To this end, the National Institute of Standards and Technology (NIST) created the Phish Scale methodology for determining how difficult a phishing email is to detect [1]. Recent research on the Phish Scale has focused on improving the robustness of the method. This paper presents initial results of the ongoing developments of the Phish Scale, including work towards the repeatability and validity of the Phish Scale using operational phishing training exercise data. Also highlighted are the ongoing efforts to minimize the ambiguities and subjectivity of the Phish Scale, as well as the design of a study aimed at gauging the usability of the scale via testing with phishing exercise training implementers.

Keywords: Usable cybersecurity · Cybersecurity awareness training · Phishing · NIST Phish Scale

1 Introduction

Over half of all emails sent and received are spam; and an ever-growing number of those messages contain malicious threats [2]. Moreover, 10% of spam messages manage to get through email filters, and phishing emails account for approximately one-third of those emails [3]. Phishing emails are malicious threats designed to deceive and extract sensitive information from the email's recipient [4]. The phishing cyber threat exploits vulnerabilities in organizations of all types and sizes, including industry, academia, and government [5–8]. A major problem with phishing is that it targets what is possibly the most vulnerable element within any security system, the human user. While spam filters are capable of filtering phishing emails based on their sender, format, and verbiage, there are still phishes which get through this net, and it is these emails which can wreak havoc on an otherwise secured system. By clicking links and volunteering personally

C. Stephanidis et al. (Eds.): HCII 2021, CCIS 1420, pp. 383–390, 2021.
https://doi.org/10.1007/978-3-030-78642-7_52

identifiable information, those who have been successfully phished can end up costing themselves and their organizations a significant amount of money in recovery efforts and time lost.

To help combat the phishing threat, organizations strive to improve phishing awareness via embedded phishing training exercises. These exercises provide organizations data – click decisions in an operational environment – from realistic, safe, and controlled training experiences. However, these click decision data that show when an email user did or did not click on a link or attachment in a phish do not tell the whole story. The National Institute of Standards and Technology (NIST) Phish Scale (NPS) was conceived to provide context to these data – click rates – and to better understand why people do or do not fall for a phish [4]. The NPS is a method for determining how difficult or easy a phishing email is to detect [1] by considering both the characteristics of the email itself and the user context of the email's recipient. Ongoing research on the use of the NPS is intended to improve its robustness, validity, and ease of use. The goal of the research presented in this paper was to assess the repeatability and validity of the NPS when applied to phishing emails used during embedded phishing awareness training exercises.

2 Applying the Phish Scale

The Phish Scale was created to provide a metric for training implementers to gain a better understanding of the variability in click rates resulting from their phishing training exercises. The output of the NPS – a difficulty rating – can be used to provide context to these click rates. Steves, et al. previously described the NPS, its development, and its components in elaborate detail [1, 9]; a high level summary is presented below.

The NPS method is comprised of two major components. The first component is a measure of the observable characteristics, or cues, of the email itself (e.g., spelling, grammar). The more cues in a phish, the easier it is to detect. The second component, the premise alignment, measures how well an email aligns with the context of one's work. The higher the premise alignment, the more difficult the phish is to detect. For example, a phish that requests payment of an invoice is more difficult to detect (high premise alignment) to an individual in the accounts payable division. While the same invoice phish might be more easily to detect to a system architect whose job duties do not include payment of invoices. The NPS includes two separate approaches to determining the premise alignment – a *Formulaic Approach* and *Blended Perspective* [9]; the former approach is the focus of the analysis in this paper. When analyzed collectively, these two NPS components produce a difficulty rating for a target audience's susceptibility to a particular phishing email (Table 1). Phishing emails with a High premise alignment and Few cues are usually harder for individuals to detect. Conversely, emails with Low premise alignment and Many cues are easier to detect by individuals.

3 Research Methodology

One of the goals of this research presented in this paper is to gauge the repeatability of applying the NPS to phishing emails. To this end, the NPS was evaluated to measure the

Table 1. Determining detection difficulty

Number of cues	Premise alignment	Detection difficulty
Few (more difficult)	High	Very difficult
	Medium	Very difficult
	Low	Moderately difficult
Some	High	Very difficult
	Medium	Moderately difficult
	Low	Moderately to Least difficult
Many (less difficult)	High	Moderately difficult
	Medium	Moderately difficult
	Low	Least difficult

agreement between independent ratings of phishing exercise emails. This effort began with the process of reevaluating the phishing emails from a previously published paper on the NPS [9]. First, a team of NIST researchers (n = 3) who were not among the original authors of the NPS independently applied the NPS to the ten phishing emails originally published by Steves, et al. [9]. The team then met to assess and compare the individual scores for both cues and premise alignment. Points of divergence in cue counts or premise alignment element scores were discussed and ultimately resolved by averaging the scores of the team members. Finally, the team's consolidated scores and difficulty rating were evaluated against the previously published findings.

Another goal of this research is to validate the metric by applying the NPS to a broader set of phishing data. The first step toward this goal is presented in this paper; the NPS was applied to three additional phishing emails used in embedded phishing awareness training exercises throughout 2020 (see Appendix). The aforementioned steps were repeated in this effort – independent ratings by research team members followed by discussion and resolution of scoring conflicts. The results of these two research efforts are presented in Sect. 4.

4 Results

This section covers the results of our analysis of applying the NPS to the ten original phishing emails as well as the three additional phishing emails. As mentioned in the previous section, these 13 emails (ten original, three new) were independently rated by members of the research team and consolidated into final scores for the cue count and premise alignment. Each email was ultimately given an overall detection difficulty rating (referred to throughout the remainder of this section as the "new" scores and ratings).

For the original ten emails, the new scores and ratings were compared to the prior published work (see Sect. 3). This comparison is detailed in Table 2, where cue and premise alignment categories are specified for each phishing email, followed by the associated numerical score in parentheses. In addition to these data, the click rates (showing the percentage of email users who clicked in the email) for the phishing exercise associated with each phish are presented in Table 2.

When comparing cues, there is clear variance in the actual scores between the new and original analysis for the individual phishing emails. However, when abstracting up to their corresponding cue categories, agreement between the new categorical data and the original categorical data was met in 90% of these phishing emails. In regard to the premise alignment, the more subjective component of the NPS, an even greater variance is seen in the actual scores given by both new and original ratings. The effects of this numerical variance can be seen in the agreement between new categories and the original categories where ratings only matched up in 40% of the phishing emails. Lastly, in large part due to the variance in the results of applying the premise alignment component, the detection difficulty ratings were agreed upon in 50% of the ten original phishing emails when comparing the original data to the new data. Given the five possible ratings on the scale of detection difficulty, it is important to note that while 50% of the new ratings did not match the original ratings, the differences were only by a factor of one (e.g., “very” to “moderately” rather than “very” to “least”). Additionally, when comparing the original detection difficulty ratings and click rates to the new ratings and corresponding click rates, the new ratings exhibit a similar pattern to the original ratings in how they line up with the click rates.

Table 2. Comparison of NIST Phish scale ratings for original phish emails

Phish email	Cues (new)	Cues (original)	Premise alignment (new)	Premise alignment (original)	Difficulty (new)	Difficulty (original)	Click rates
E1	Few (6)	Few (7)	Low (10)	High (30)	Moderate	Very	49.3%
E2	Some (10)	Some (14)	Medium (13)	High (24)	Moderately	Very	43.8%
E3	Few (7)	Few (8)	Medium (16)	High (24)	Very	Very	20.5%
E4	Some (9)	Few (6)	Medium (14)	High (18)	Moderately	Very	19.4%
E5	Some (9)	Some (11)	Low (9)	Medium (14)	Moderately to least	Moderately	11.6%
E6	Some (13)	Some (13)	Low (0)	Low (10)	Moderately to least	Moderately to least	11.0%
E7	Many (18)	Many (18)	Medium (13)	Medium (16)	Moderately	Moderately	9.1%
E8	Some (9)	Some (12)	Medium (12)	Medium (12)	Moderately	Moderately	8.7%
E9	Some (14)	Some (11)	Low (−1)	Low (2)	Moderately to least	Moderately to least	4.8%
E10	Some (10)	Some (12)	Medium (13)	Low (4)	Moderately	Moderately to least	3.2%

Table 3 features the averaged calculations for the three independent raters of the current study. The click rates for emails E11 and E12 align well with their respective detection difficulty ratings, according to the pattern exhibited in the application of the NPS to the previous ten emails. However, the click rates for email E13 do not fully align with the established detection difficulty rating scale. The trend of the NPS has been for emails with a click rate as low as 2.8% to have a “least” or “moderately to least” difficulty rating.

Table 3. NIST Phish Scale ratings for new phish emails

Phish emails	Cues	Premise alignment	Detection difficulty	Click rates
E11	Some (12)	Low (4)	Moderately to least	12.7%
E12	Many (18)	Low (9)	Least	5.4%
E13	Many (16)	Medium (13)	Moderately	2.8%

The NPS has the ability to contextualize click rates with its detection difficulty ratings. However, there are some unexpected factors which may inflate click rates which could lead to the disagreement between click rates and detection difficulty ratings (as exhibited by E13). For example, when a phishing email appeared to come directly from an authority figure in upper management, it elicited serious concerns and a deeper sense of action by the email recipient than was measurable by the NPS, ultimately leading to an increased click rate and the aforementioned disagreement. These factors are intended to be addressed in future iterations of NPS development.

5 Discussion and Future Work

This paper presents an initial look into the ongoing validation effort of the NPS. The results discussed in the previous section show the margin of error in the NPS difficulty rating determination; there can be a slight variance in independent scores of a phishing email, yet that variance is not reflected in the resulting detection difficulty rating. This provides insight into the development of future iterations of the NPS; however, additional validation testing is needed, including testing with larger and more diverse datasets. To this end, the NPS is currently being tested with a variety of large datasets (both public and nonpublic) from universities, private companies, and other government agencies. The findings from applying the NPS to a variety of datasets will be used to improve future iterations of the NPS. These efforts are aimed at ensuring the NPS's accuracy and validity.

NIST is conducting a research study to determine the usability and applicability of the NPS. The study invited both federal and non-federal organizations with robust phishing programs to apply the NPS in their organizations, aligning with their existing embedded phishing awareness training programs. Following their use of the NPS, training implementers were asked to provide detailed feedback and recommendations about their use of the NPS. This valuable real-world information resulting from the study will determine the effectiveness of the NPS in unique organizational environments, how usable the NPS is, and how organizations use the NPS to contextualize phishing exercise click rates.

As mentioned throughout this paper, NPS research is ongoing. Current efforts to improve repeatability, to evaluate validity, and to assess the usability of the NPS are expected to lead to a more streamlined version of the NPS that would be beneficial to organizations to provide clarity, functionality, and adaptability of the metric. Future iterations of the NPS will incorporate various modifications grounded in findings from the research. Revisions currently being considered for adoption and inclusion are: 1) for the observable cues component, reducing subjectivity, increasing identification accuracy, and minimizing redundancy across the scale, 2) refining the cue counting method by incorporating a weighting metric to address cue saliency, and 3) restructuring the premise alignment's five elements to be more efficient, reducing the total number of elements and adopting proven methodologies for determination of premise alignment element scores. Additionally, insights gleaned from the aforementioned usability study, including the identification of successful practices and strategies, and lessons learned will be used to refine future iterations of the NPS.

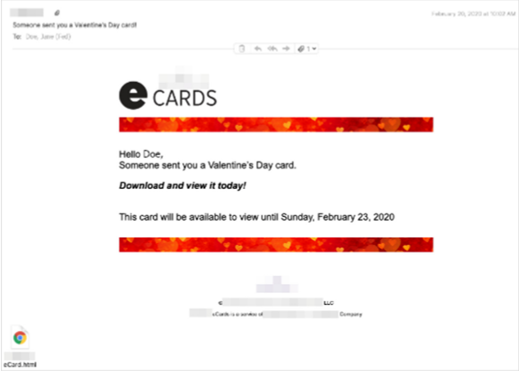
6 Conclusion

The NPS helps organizations and phishing awareness training implementors in two primary ways. Firstly, by contextualizing message click and reporting rates for a target audience, and secondly by providing a way to characterize actual phishing threats so training implementors can reduce the organization's security risk. Organizations should tailor their cybersecurity and privacy awareness training program to their unique environment while still meeting their organizations' mission and risk tolerance. Likewise, the NPS goes beyond the face value of an email by accounting for the environment, roles, and responsibilities of people within an organization. Tailoring training to the types of threats their organization faces helps them maintain a resilient security and privacy posture. Additionally, when click rates and quantitative and qualitative metrics from the NPS are viewed holistically, they can signal to an organization that training approaches and objectives, delivery methods, training frequency or content necessitate alterations to be effective in combating the ever-changing phishing threat landscape.

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.

Appendix

Note: Logos have been blinded from the phishing email images below (Figs. 1, 2 and 3).



[redacted]

Someone sent you a Valentine's Day card!

To: Doe, Jane (Fed)

February 20, 2020 at 10:02 AM

Hello Doe,
Someone sent you a Valentine's Day card.

Download and view it today!

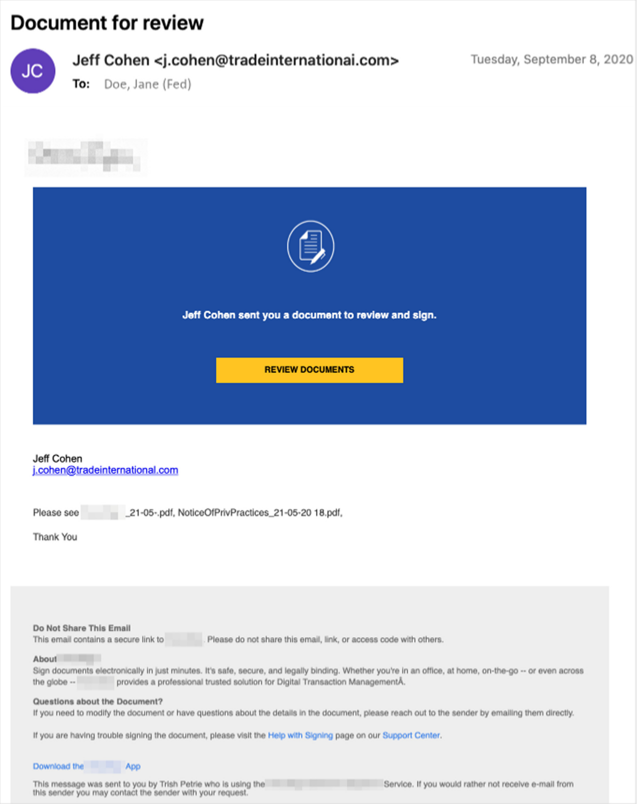
This card will be available to view until Sunday, February 23, 2020

[redacted]

© [redacted] LLC

[redacted] eCards is a service of [redacted] Company

Fig. 1. E11: E-card phish



Document for review

Jeff Cohen <j.cohen@tradeinternationai.com>

To: Doe, Jane (Fed)

Tuesday, September 8, 2020

[redacted]

Jeff Cohen sent you a document to review and sign.

Jeff Cohen
j.cohen@tradeinternational.com

Please see [redacted] _21-05-.pdf, NoticeOfPrivPractices_21-05-20 18.pdf,

Thank You

Do Not Share This Email
This email contains a secure link to [redacted]. Please do not share this email, link, or access code with others.

About [redacted]
Sign documents electronically in just minutes. It's safe, secure, and legally binding. Whether you're in an office, at home, on-the-go -- or even across the globe -- [redacted] provides a professional trusted solution for Digital Transaction ManagementÂ.

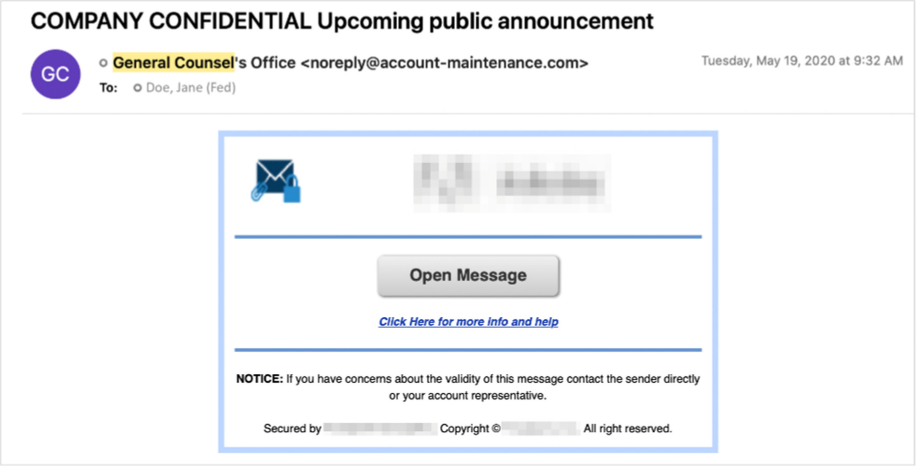
Questions about the Document?
If you need to modify the document or have questions about the details in the document, please reach out to the sender by emailing them directly.

If you are having trouble signing the document, please visit the Help with Signing page on our Support Center.

Download the [redacted] App

This message was sent to you by Trish Petrie who is using the [redacted] Service. If you would rather not receive e-mail from this sender you may contact the sender with your request.

Fig. 2. E12: Document review phish



COMPANY CONFIDENTIAL Upcoming public announcement

GC General Counsel's Office <noreply@account-maintenance.com> Tuesday, May 19, 2020 at 9:32 AM

To: Doe, Jane (Fed)

[redacted]

NOTICE: If you have concerns about the validity of this message contact the sender directly or your account representative.

Secured by [redacted] Copyright © [redacted] All right reserved.

Fig. 3. E13: Public Announcement phish

References

1. Steves, M.P., Greene, K.K., Theofanos, M.F.: A phish scale: rating human phishing message detection difficulty. In: Proceedings 2019 Workshop on Usable Security. Workshop on Usable Security, San Diego, CA. (2019). https://doi.org/10.14722/usec.2019.23028
2. Ezpeleta, E., Velez de Mendizabal, I., Hidalgo, J. M.G., Zurutuza, U.: Novel email spam detection method using sentiment analysis and personality recognition. Logic J. IGPL, **28**(1), 83–94 (2020). https://doi.org/10.1093/jigpal/jzz073
3. Cyren: Email Security Gap Analysis: Aggregated Results. (2017). https://pages.cyren.com/rs/944-PGO-076/images/Cyren_Report_GapAnalysisAgg_201711.pdf. Accessed Jan 2021
4. Greene, K.K., Steves, M., Theofanos, M., Kostick, J.: User context: an explanatory variable in phishing susceptibility. In: Proceedings of 2018 Workshop Usable Security (USEC) at the Network and Distributed Systems Security (NDSS) Symposium (2018)
5. Aaron, G., Rasmussen, R.: Global phishing survey 2016: Trends and domain name use. Anti-Phishing Working Group. June 2017. https://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf. Accessed Aug 2017
6. Hong, J.: The state of phishing attacks. Commun. ACM **55**(1), 74–81 (2012)
7. SonicWall: 2019 SonicWall Cyber Threat Report (2019). https://www.sonicwall.com/resources/white-papers/2019-sonicwall-cyber-threat-report. Accessed Aug 2020
8. Symantec: Internet Security Threat Report. vol. 24 (2019). https://www.symantec.com/security-center/threat-report. Accessed Aug 2020
9. Steves, M., Greene, K., Theofanos, M.: Categorizing human phishing difficulty: a Phish scale. J. Cybersec. **6**(1), tyaa009 (2020). https://doi.org/10.1093/cybsec/tyaa009

NIST SP 1283
August 2022

Sub Topic: Fire Research

12^{th} U.S. National Combustion Meeting
Organized by the Central States Section of the Combustion Institute
May 24-26, 2021
College Station, Texas

A second-generation phi meter for global equivalence ratio and gas species concentration measurements

Ryan Falkenstein-Smith[1*] *and Thomas Cleary*[1]

[1]*National Institute of Standards and Technology, Gaithersburg, MD, USA*
[*]*Corresponding author:* ryan.falkenstein-smith@nist.gov

Abstract: This work presents a second-generation phi meter capable of making simultaneous real-time measurements of the global equivalence ratio and combustion product concentrations (i.e., water vapor, carbon dioxide, and oxygen). The fuels initially examined are methane, propane, and propene. Fuel is mixed with predetermined air concentrations to create a broad spectrum of fuel-air ratios. The global equivalence ratio of a sampled mixture is measured by the phi meter via lean combustion. Water vapor concentrations are made using a thermoelectric cooler and supplemental drying unit placed between a high-temperature mass flow controller and ambient temperature mass flow meter. The difference between the flow units provides the water vapor concentration measurement. Carbon dioxide and oxygen concentration measurements are made using IR and paramagnetic sensors, respectively. The ability to measure combustion product concentrations provides additional insight into the sampled gas, specifically, carbon to hydrogen ratio calculations and total mass balance. The measured and calculated global equivalence ratios for all gas mixtures are observed to be in fair agreement. The calculated values and estimated water vapor concentration measurements are consistent within the experimental uncertainty. The uncertainty of the calculated carbon to hydrogen ratio of incoming gas mixtures is within the parent fuels' ratio, thus validating the instrument's ability to measure gas species concentrations of the combustion products.

Keywords: Gas extractive sampling, Lean combustion, Mass balance, Phi meter, Equivalence ratio

1. Introduction

Ventilation can have a significant impact on the dynamics of an enclosure fire. The availability of oxygen within an enclosure is known to affect the generation of gas species, particularly CO, which can increase the threat posed by fire. Ventilation can be defined from the equivalence ratio within a compartment, either on a global or local scale. In a compartment of known volume, the global equivalence ratio is defined by the total mass of air entering and the total mass of fuel occupying said compartment. The local equivalence ratio is defined similarly, but only by the mass of air and fuel within a portion of the total volume of the compartment. In relation to each other, the global equivalence ratio of a given volume is equal to sum of the local equivalence ratios throughout said volume. Further discussion on the global equivalence ratio concept is provided in Refs. [1, 2].

There have been several techniques used to measure the equivalence ratio from combustion processes. Applications of molecular spectroscopy, such as laser induced fluorescence (LIF) and laser induced breakdown spectroscopy (LIBS), are shown to measure the equivalence ratio using a non-extractive technique [3–7] and require extensive corrections to address low signal to noise

Falkenstein-Smith, Ryan; Cleary, Thomas. "A second-generation phi meter for global equivalence ratio and gas species concentration measurements." Presented at 12th US National Combustion Meeting (12th US NCM). May 24, 2021 - May 26, 2021.

ratios. Gas chromotography with mass selectivity detectors can provide time-averaged equivalence ratio using an extractive sampling approach [8–11]. For real-time applications, a phi meter, originally proposed by Babrauskas et al. [12], has been utilized for several applications [13] to make in-situ global equivalence ratio measurements.

The purpose of this study is to examine a second generation phi meter capable of measuring the global equivalence ratio simultaneously with the total mass in the system, via O_2, CO_2, and H_2O concentration measurements. A series of combustion experiments are performed to demonstrate the capability of the enhanced phi meter. Methane, propane, and propene are selected as the fuels of interest. Predetermined ratios of fuel and air spanning from fuel-rich to fuel-lean conditions are examined.

2. Methods and Experimental Approach

Figure 1 presents a schematic of the second-generation phi meter. Extracted gas samples are driven through the phi meter by a vacuum pump positioned at the end of the sample line. Upon extraction, gas samples are fed into the phi meter's reactor component enclosed in a high-temperature tubular furnace. The reactor component includes quartz tubing packed with a combustion catalyst, heated to approximately 900 °C. In this work, platinum-coated silica beads (Sigma-Aldrich/Millipore Sigma 520691[1]) are selected based on their high performance in previous phi meter designs [14].

Excess oxygen is introduced to the extracted gas sample at the inlet of the phi meter via an inner tube within the quartz tubing. The addition of oxygen to the incoming gas mixture, combined with the presence of a high-temperature combustion catalysts, allows for a lean combustion with a reactor exhaust exclusively comprised of O_2, CO_2, H_2O, and N_2. Upon exiting the reactor, the exhaust flow is fed through a high-temperature mass flow controller (Alicat MCRW-5SLPM-D-HT/5M). The high-temperature mass flow controller regulates the total mass flow moving through the reactor, which is preset for each experiment. The high-temperature mass flow controller is heated using a 24 VDC heating element maintained at approximately 90 °C. The elevated temperature of the high-temperature mass flow controller prevents water vapor from the reactor's exhaust condensing, thus allowing the controller to account for the vapor mass in the sample line.

A thermoelectric cooler, condensate reservoir, and supplemental drying unit that incorporates Nafion™ tubing are sequentially positioned behind the high-temperature mass flow controller for the purpose of scrubbing water from the sample line. Once dried, the sample gas flows through a mass flow meter where the mass flow of the dried gas is measured. A portion of the mass flow meter's output is sampled into a Servoflex 5200 MiniMP Gas Analyzer, equipped with paramagnetic and infrared sensors calibrated to measure O_2 and CO_2 volume fractions.

The global equivalence ratio, ϕ_{G}, can be calculated by the phi meter's O_2 and CO_2 volume fraction, $X_{\mathrm{O_2,A}}$ and $X_{\mathrm{CO_2,A}}$, and mass flow meter volumetric flow, $\dot{V}_{\mathrm{MFM}}$, measurements in the equation below:

$$\phi_{\mathrm{G}} = 1 + \left(\frac{1 - X_{\mathrm{O_2,Ent}}}{X_{\mathrm{O_2,Ent}}(1 - X_{\mathrm{O_2,A}} - X_{\mathrm{CO_2,A}})}\right)\left(\frac{\dot{V}_{\mathrm{O_2,Ex}}}{\dot{V}_{\mathrm{MFM}}} - X_{\mathrm{O_2,A}}\right) \tag{1}$$

[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.

Sub Topic: Fire Research

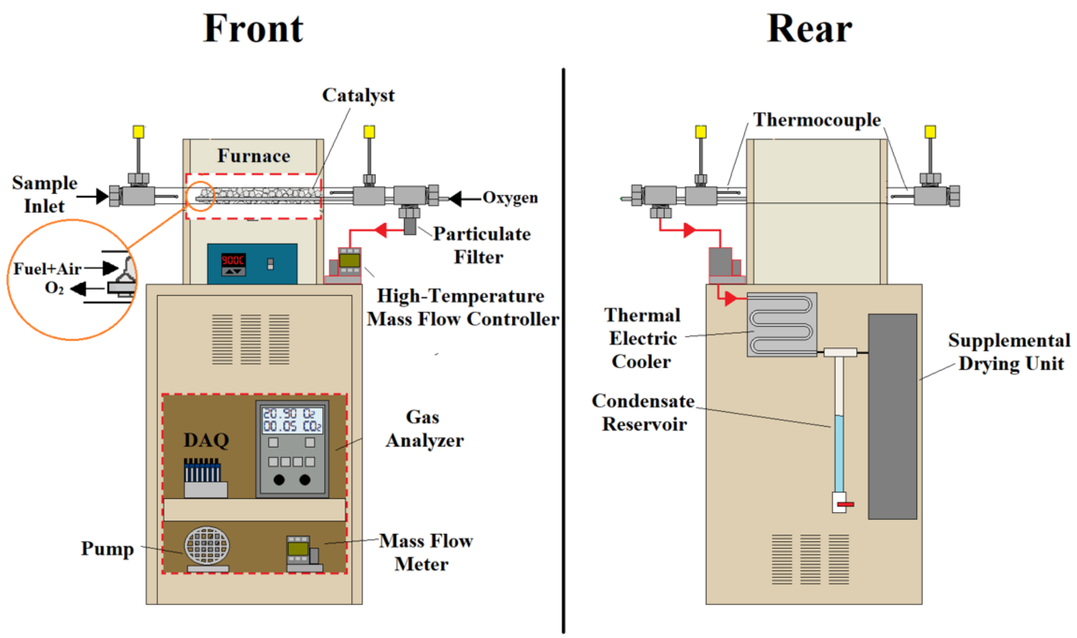


Figure 1: Schematic of the second-generation phi meter

where $X_{\mathrm{O_2,Ent}}$ is the volume fraction of oxygen in the air (approx. 20.95%) and $\dot{V}_{\mathrm{O_2,Ex}}$ is the volumetric flows of the excess oxygen, respectively. A full dry basis derivation of Eq. 1 is documented in Ref. [12].

The portion of water vapor in the exhaust flow from the reactor, $\dot{V}_{\mathrm{H_2O}}$, is measured from the difference between the volumetric flow readings at the high-temperature mass flow controller, $\dot{V}_{\mathrm{HTMFC}}$, and mass flow meter.

$$\dot{V}_{\mathrm{H_2O}} = \dot{V}_{\mathrm{HTMFC}} - \dot{V}_{\mathrm{MFM}} \tag{2}$$

This calculation assumes that the thermoelectric cooler, condensate reservoir, and supplemental drying unit are completely drying the reactor's exhaust and that no condensate is within the sample line at the mass flow meter.

As a way to verify the accuracy of the experimental design, the carbon to hydrogen ratio is calculated from the CO_2 and H_2O volume fractions, $X_{\mathrm{CO_2}}$ and $X_{\mathrm{H_2O}}$, determined at the high-temperature mass flow controller. This calculation is represented in Eq. 3 shown below:

$$\frac{\mathrm{C}}{\mathrm{H}} = \frac{\mathrm{x}_i X_{\mathrm{CO_2}}}{\mathrm{y}_i X_{\mathrm{H_2O}}} \tag{3}$$

Here the C/H ratio is determined from the CO_2 and H_2O at the high-temperature mass flow controller, since all fuel entering the phi meter is combusted. The number of carbon and hydrogen atoms in the molecule are represented by x_i and y_i, respectively. The carbon to hydrogen ratio of the fuel molecules examined in this work are reported in Table 1 and shown in Fig 4.

Falkenstein-Smith, Ryan; Cleary, Thomas. "A second-generation phi meter for global equivalence ratio and gas species concentration measurements." Presented at 12th US National Combustion Meeting (12th US NCM). May 24, 2021 - May 26, 2021.

Table 1: List of carbon to hydrogen (C/H) ratio of fuels

Fuel	Methane	Propane	Propene
C/H	1/4	3/8	1/2

Incoming gas mixtures comprised of fuel and air are controlled via the phi meter inlet flow. The inlet to the phi meter is partitioned such that fuel flow maintained by an external mass flow controller mixed with ambient air are sampled into the phi meter. The incoming air flow is measured by an additional mass flow meter. The fuel-air ratio of the incoming mixture is determined from controlled fuel flow and measured air flow, which is determined by the difference between the total mass flow set by the high-temperature mass flow controller and the mass flow of the fuel.

A mass balance evaluation is applied to provide a more in-depth verification method. The total mass flow present at the inlet of phi meter, $\dot{m}_{\mathrm{in}}$, which includes the mass of incoming fuel, $\dot{m}_{\mathrm{Fuel}}$, air, $\dot{m}_{\mathrm{Air}}$ and excess oxygen, $\dot{m}_{\mathrm{O_2,Ex}}$, is compared to the total mass flow at the high-temperature mass flow controller, $m_{\mathrm{H\dot{T}MFC}}$, comprised of the mass of remaining oxygen, $\dot{m}_{\mathrm{O2,HTMFC}}$, carbon dioxide, $\dot{m}_{\mathrm{CO_2,HTMFC}}$, water vapor, $\dot{m}_{\mathrm{H_2O,HTMFC}}$, and nitrogen, $\dot{m}_{\mathrm{N_2,HTMFC}}$. Assuming no mass losses between the flow units, the mass flow of nitrogen is determined from the difference between the total mass flow measured at the mass flow meter, $\dot{m}_{\mathrm{MFM}}$, and the mass flow of O_2 and CO_2. The mass flow of O_2 and CO_2 is calculated from the product of the gas analyzer O_2 and CO_2 volume fraction measurements and the total mass flow meter reading, as shown in Eq. 7.

$$\dot{m}_{\mathrm{in}} = \dot{m}_{\mathrm{HTMFC}} \tag{4}$$

$$\dot{m}_{\mathrm{Fuel}} + \dot{m}_{\mathrm{Air}} + \dot{m}_{\mathrm{O_2,Ex}} = \dot{m}_{\mathrm{O_2,HTMFC}} + \dot{m}_{\mathrm{CO_2,HTMFC}} + \dot{m}_{\mathrm{H_2O,HTMFC}} + \dot{m}_{\mathrm{N_2,HTMFC}} \tag{5}$$

where

$$\dot{m}_{\mathrm{N_2,HTMFC}} = \dot{m}_{\mathrm{MFM}} - \dot{m}_{\mathrm{O_2,HTMFC}} - \dot{m}_{\mathrm{CO_2,HTMFC}} \tag{6}$$

or

$$\dot{m}_{\mathrm{N_2,HTMFC}} = \dot{m}_{\mathrm{MFM}}(1 - X_{\mathrm{O_2}} - X_{\mathrm{CO_2}}) \tag{7}$$

All measurements are collected and monitored by a data acquisition system. Data samples are recorded at 1 Hz for a 1 minute period. All experimental conditions are repeated at least twice. An extensive uncertainty analysis of all measurements is provided in Ref. [15]. Measurement uncertainties are estimated using the law of propagation of uncertainty which combines the Type A and B evaluation of uncertainty. For most measurements, the Type B evaluation of uncertainty is the dominant error. Unless otherwise stated, the uncertainty of the measurements are expressed assuming a 95% confidence level.

3. Results and Discussion

Figure 2 shows the comparison between global equivalence ratio measurements determined from Eq. 1 and the predetermined fuel-air mixtures at the inlet. The dotted line represents equivalency between the calculated and measured values. Since the composition of the sampled flow is consistent, the local and global equivalence ratios are equivalent. The measured ϕ_{G} for all fuel and fuel-air ratios are shown to be in agreement with the monitored incoming gas mixtures. The consistency between the phi meter measurements and estimated incoming equivalence ratio suggest that the incoming fuel is completely combusted within the reactor under the given configuration (i.e., reactor setup and high-temperature mass flow controller and furnace temperature settings). The indication that all incoming fuel is combusted within the phi meter suggest that the calculated water vapor flow is half the product of the number of hydrogen atoms in the incoming fuel, y and the volumetric flow of the fuel, $\dot{V}_{\mathrm{Fuel}}$, as represented below:

$$\dot{V}_{\mathrm{H_2O,HTMFC}} = \frac{y}{2}\dot{V}_{\mathrm{Fuel}} \tag{8}$$

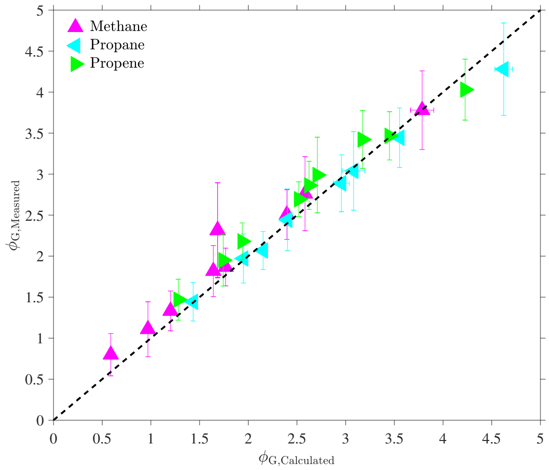


Figure 2: Measured global equivalence ratio via Eq. 1 as a function of the global equivalence ratio estimated from predetermined fuel-air mixtures. The dotted line represents the equivalency between the calculated and measured values.

The comparison between the measured and calculated water vapor flow using Eq. 2 and 8, respectively, is presented in Fig 3. The water vapor flow measurements determined from the difference in flow readings are observed to match the calculated water vapor flows within the experimental uncertainty, indicated by the dotted line where the calculated and measured values are equivalent. In the cases of propane, a discrepancy between the calculated and measured flow appears to increase when water vapor concentration is greater than 15 %. The deviation between the calculated and measured water vapor flows at high concentrations may be attributed to the inadequate performance of the thermoelectric cooler or supplemental drying unit. An increase in the flow's residence time in the thermoelectric cooler and supplemental dryer or a decrease in the thermoelectric coolers operating temperature, may reduce the discrepancy between the measured and calculated values.

The estimated carbon to hydrogen ratio from the parent fuel, indicated by the dotted line, and the calculated ratio determined from the CO_2 and H_2O measurements are plotted as a function of the measured ϕ_{G} and shown in Fig. 4. For all measurements, the experimental uncertainty of the calculated C/H is within parent fuels' ratios. In several instances, the theoretical and calculated values are nearly matching, verifying the accuracy of the CO_2 and H_2O concentration measurements. The agreement of the theoretical and measured carbon to hydrogen ratios demonstrates the

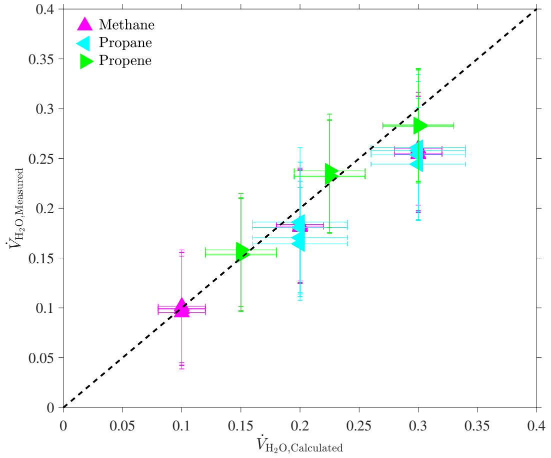


Figure 3: Volumetric flow of water vapor via Eq. 2 as a function of the anticipated volumetric flow of water vapor estimated from Eq. 8. The dotted line represents the equivalency between the calculated and measured values.

enhanced phi meter's potential for investigating combustion processes by measuring the C/H ratio without any knowledge of the parent fuel.

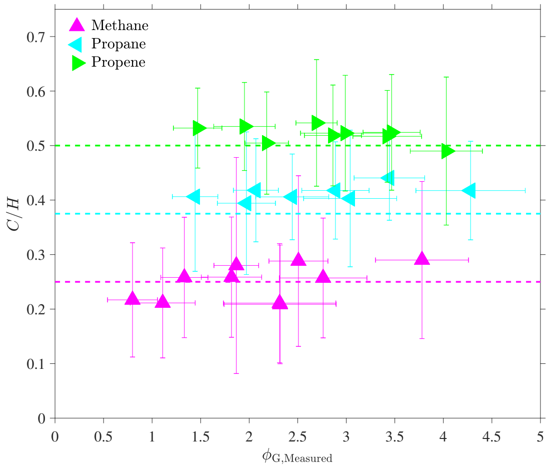


Figure 4: Carbon to hydrogen ratio calculated from the CO_2 and H_2O species measurements compared to the theoretical values (dotted lines) as a function of the measured global equivalence ratio.

Figure 5 shows the ratio between the total mass calculated at the high temperature mass flow controller and measured at the inlet of the phi meter as a function of the measured ϕ_G. Unity is represented by a dotted line. For each measurement, unity falls within the experimental uncertainty of each mass ratio. In some cases, the calculated mass ratios are near unity, which indicate minimal mass loss within the sample line. The agreement between masses for varying equivalence ratios and fuels further supports the validity of the O_2, CO_2, H_2O, and N_2 concentration measurements and more importantly the design of the phi meter.

Sub Topic: Fire Research

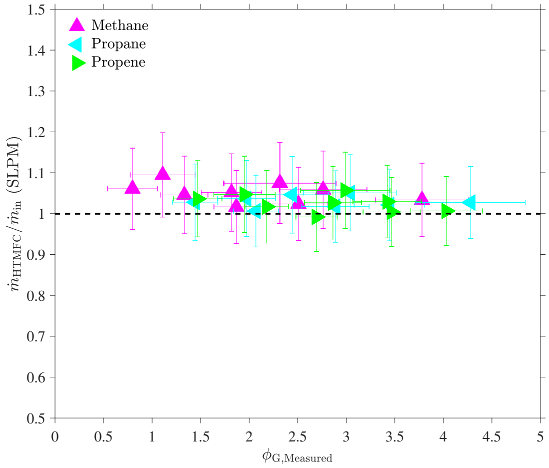


Figure 5: The ratio of the mass flow rate at the high temperature mass flow controller to the inlet of the phi meter calculated via Eqs.4-6 as a function of the measured global equivalence ratio. The dotted line represents the equivalency between the calculated and measured values.

4. Conclusions

This study demonstrates the capabilities of a second-generation phi meter. Global equivalence ratios, water vapor concentrations, carbon to hydrogen ratios, and total mass flow are measured and compared to calculated values. Comparisons between the measured and calculated values are observed to be in fair agreement with the calculated values, within experimental uncertainty. The enhanced phi meter's capability to measure the global equivalence ratio simultaneously with O_2, CO_2, and H_2O has significant potential for investigating combustion processes. For example, before entering the phi meter a portion of an extracted gas sample can be partitioned into an additional gas analyzer capable of measuring O_2, CO_2, and CO. From a combination of the additional gas analyzer and phi meter measurements, in-situ gas species concentrations measurements and the local equivalence ratio can be determined without any knowledge of the parent fuel. Future work will focus on demonstrating this approach in large-scale applications.

References

[1] W. Pitts, The global equivalence ratio concept and the formation mechanisms of carbon monoxide in enclosure fires, Progress in energy and combustion science 21 (1995) 197–237.

[2] C. Wieczorek, U. Vandsburger, and J. Floyd, An evaluation of the global equivalence ratio concept for compartment fires: data analysis methods, Journal of Fire Protection Engineering 14 (2004) 9–31.

[3] P. Stavropoulos, A. Michalakou, G. Skevis, and S. Couris, Quantitative local equivalence ratio determination in laminar premixed methane–air flames by laser induced breakdown spectroscopy (LIBS), Chemical Physics Letters 404 (2005) 309–314.

Sub Topic: Fire Research

[4] B. McGann, T. Ombrello, D. Peterson, E. Hassan, S. Hammack, C. Carter, T. Lee, and H. Do, Lean fuel detection with nanosecond-gated laser-induced breakdown spectroscopy, Combustion and Flame (2021).

[5] S. Zhang, X. Yu, F. Li, G. Kang, L. Chen, and X. Zhang, Laser induced breakdown spectroscopy for local equivalence ratio measurement of kerosene/air mixture at elevated pressure, Optics and Lasers in Engineering 50 (2012) 877–882.

[6] M. Tripathi, K. Srinivasan, S. Krishnan, F. Yueh, and J. Singh, A comparison of multivariate LIBS and chemiluminescence-based local equivalence ratio measurements in premixed atmospheric methane–air flames, Fuel 106 (2013) 318–326.

[7] D. Han and R. Steeper, An LIF equivalence ratio imaging technique for multicomponent fuels in an IC engine, Proceedings of the Combustion Institute 29 (2002) 727–734.

[8] R. Falkenstein-Smith, K. Sung, J. Chen, K. Harris, and A. Hamins, The Structure of Medium-Scale Pool Fires, Second Edition, NIST Technical Note Report No. 2082e2, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2021.

[9] R. Falkenstein-Smith, K. Sung, J. Chen, and A. Hamins, The Chemical Structure of a 30 cm methanol pool fire, Fire and Materials 45 (2021) 429–434.

[10] R. Falkenstein-Smith, K. Sung, J. Chen, and A. Hamins, Chemical Structure of Medium-Scale Liquid Pool Fire, Fire Safety Journal (2020) 103099.

[11] R. Falkenstein-Smith, K. Harris, K. Sung, T. Liang, and A. Hamins, A calibration and sampling technique for quantifying the chemical structure in fires using GC/MSD analysis, Fire and Materials (2021).

[12] V. Babrauskas, W. Parker, G. Mulholland, and W. Twilley, The phi meter: A simple, fuel-independent instrument for monitoring combustion equivalence ratio, Review of scientific instruments 65 (1994) 2367–2375.

[13] B. Andersson, V. Babrauskas, G. Holmstedt, S. Särdqvist, and G. Winter, Scaling of combustion products: Initial results from the TOXFIRE study, Industrial Fires III Workshop Proceedings (1997), pp. 65–74.

[14] B. Andersson, G. Holmstedt, and A. Dagneryd, Determination of the equivalence ratio during fire, comparison of techniques, Fire Safety Science 7 (2003) 295–308.

[15] R. Falkenstein-Smith and T. Cleary, The Design and Performance of a Second-Generation Phi Meter, NIST Technical Note Report No. In Progress, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2022.

Sub Topic: Fire Research

12^{th} U.S. National Combustion Meeting
Organized by the Central States Section of the Combustion Institute
May 24-26, 2021
College Station, Texas

The structure of medium-scale propane pool fires

Ryan Falkenstein-Smith[*1], *Kunhyuk Sung*[1], *and Anthony Hamins*[1]

[1]*National Institute of Standards and Technology, Gaithersburg, MD, USA*
[*]*Corresponding author:* ryan.falkenstein-smith@nist.gov

Abstract: A series of time-averaged temperature, velocity, and gas species measurements are made along the centerline of propane fires established on a 37 cm diameter gas burner situated in a quiescent environment. Fires with heat release rates of 20 kW and 34 kW are selected for study to complement previous measurements of total heat flux emitted to the surroundings in these fires. Gas samples are extracted at various heights above the burner centerline and analyzed using a Gas Chromatograph with mass selectivity and thermal conductivity detectors. Soot mass fractions are gravimetrically measured. Major gas species, including propane, oxygen, carbon dioxide, water, carbon monoxide, hydrogen, nitrogen, argon, and soot, are detected and quantified. Intermediate gas species are observed to achieve peak concentrations a few cm above the fuel surface, whereas carbon dioxide and water peak further downstream. The chemical and physical structures of the fires are compared by considering the profiles of measured gas species volume fractions, soot mass fractions, temperature and velocity profiles. Plotting the temperature and velocity profiles as a function of z* collapse the experimental data and provide insight into the structure of these medium-scale propane pool fires.

***Keywords:* Gas pool fires, Gas species measurements, Propane flame structure**

1. Introduction

Computation fluid dynamics fire models are an important component to performance based design in fire protection engineering. A requirement of their acceptance and prominent usage are verifying and validating the developed models, the latter of which involves comparison with experimental measurements. This objective of this work is to provide measurements for validation in fire model validation.

Pool fires are a convenient test bed for model validation due to their flat, horizontal, and isothermal fuel surface which provides a well-defined boundary-condition for modeling. A zone of particular interest is the fuel-rich core between the flame and pool surface, where gas species can absorb energy that would otherwise have been transferred to the fuel surface. Therefore, gas species concentrations, temperature, and velocity measurements are of interest. However, few studies in literature have reported local chemical species measurements within the flame [1–14].

The purpose of this study is to characterize the spatial distribution of the temperature, velocity, and principal chemical species in medium-scale propane pool fires steadily burning in a well ventilated, quiescent environment. Propane is selected as the fuel of interest, burned at different fire sizes. A series of fire experiments are conducted using a 37 cm effective diameter pool burner. Gaseous species, soot, temperature, and velocity measurement are made at various locations with the centerline of the propane fires. The measurement techniques applied in this study are justified based on previous work in fire literature [3, 15].

Falkenstein-Smith, Ryan; Sung, Kunhyuk; Hamins, Anthony. "The structure of medium-scale propane pool fires." Presented at 12th US National Combustion Meeting (12th US NCM). May 24, 2021 - May 26, 2021.

Sub Topic: Fire Research

2. Methods/Experimental

2.1 Pool Burner Setup

All experiments are conducted under a canopy hood surrounded by four 2.5±0.1 m x 2.5±0.1 m wire-mesh screen (nominally 5 mesh/cm), which reduces the impact of room ventilation. All measurements are made once the mass rate of the fire reaches steady-state, achieved approximately 2 min after ignition. The propane fires are burned using a round 37 cm effective diameter, porous metal burner. The gas burner is maintained at a constant temperature by circulating cooled water (approximately 20 °C ± 5 °C. Fuel to the gas burner is controlled via a Brooks mass flow controller, Model 5863[1] located outside the enclosure. Description of previous pool fires experiments are found in Refs. [2, 7, 16–19]

The mean flame height is estimated from 3600 frames of high-resolution video of the steady-state propane pool fires using MATLAB's Image Processing Toolbox. The flame height of a single frame is defined by the distance between the pool surface and observed flame tip. All measurements are repeated, then averaged to provide the mean flame height.

2.2 Temperature Measurements

Time-averaged temperature measurements are made at varying positions along the centerline profiles of the pool fires. An S-type (Pt, 10% Rh/Pt), bare-wire, thermocouple (OMEGA P10R-001) with a 50 μm wire diameter and a bead diameter approximately 150 μm is used. Temperature measurements are sampled at nominally 250 Hz for 2 min, or approximately 300 pulsing cycles [8]. Optical microscope observations show that the thermocouple bead is spherical.

To account for the unsteady flow field, thermocouple temperature measurements are corrected for heat losses and thermal inertia following Shaddix [20], in which conduction losses are assumed to be small and the compensation for thermal inertia is calculated from the Nusselt number via the Ranz-Marshall model [21]. The temperature-dependent gas properties for the Reynolds and Prandtl numbers are taken as those of air from Ref. [22]. The temperature-dependent emissivity and thermophysical properties of platinum are taken from Refs. [23, 24]. A detailed description of the measurement and its uncertainty is described in Refs. [25].

2.3 Velocity Measurements

Time-averaged velocity measurements are made using a bi-directional probe positioned about the burner's centerline. A full description of the measurement and its uncertainty is provided in Ref. [26]. The instantaneous gas velocity is determined from the pressure difference between the front and rear of the probe. The differential pressure is measured with multiple pressure transducers, with varying instrument response times, sampling between 250 to 500 Hz for a period of 2 min. Temperature-dependent gas properties are taken as those of air and are calculated using temperature measurements made approximately 5 mm below the probe. Temperature measurements are made using the same method described in Section 2.2.

[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.

2.4 Volume Fraction of Gas Species Measurements

Time-averaged gas species volume fraction, $\bar{X}_i$, measurements are made using an Agilent 5977E Series Gas Chromatograph with thermal conductivity and mass selectivity detectors (GC/MSD). Gas samples are extracted using a thermal quenching probe composed of concentric, stainless-steel tubes with an outer annular coolant flow and inner extracted sample flow. The quenching probe's temperate is maintained by a heated water reservoir (approximately 90 °C) which feeds into and out of the probe for the duration of the experiment. Directly behind the quenching probe is a heated sampling line, heated to approximately 140 °C to prevent water from condensing within the line. The heated sampling line also includes a soot filter, used to provide time-average soot measurements, and a nominal 150 ml mixing chamber, that ensures the sample is well-mixed and is an accurate representation of the local volume of interest.

The gas sampling flow rate is controlled via a mass flow controller located between the GC/MSD and vacuum pump. During the gas sampling procedure, the volumetric flow is approximately 200 mL/min and recorded at 2 Hz. Gas sampling time is dependent on the probe location within the fire, with the sampling period varying from 12 to 25 min. All gas species measurements at a given location are repeated at least twice. The mean mass fraction, $\bar{Y}_i$, of a given species i is calculated from the measured volume fraction using the following expression:

$$\bar{Y}_i = \frac{\bar{X}_i \, \mathrm{W}_i}{\sum \bar{X}_i \, \mathrm{W}_i} \tag{1}$$

where W_i is the molecular weight of a given species. A detailed description of the gas species measurement and its uncertainty is documented in several other works [8].

2.5 Soot Mass Fraction Measurements

Soot mass fraction, Y_s, is measured using a well established gravimetric technique. Soot is filtered out from the extracted gas sample directly behind the thermal quenching probe. Before sampling, a desiccated 47 mm polytetrafluoroethylene (PTFE) filter is weighed and placed into a stainless steel particulate filter holder (PALL 2220). After sampling, the PTFE filter is removed and dried in a desiccator for 48 h then weighed. After most experiments, soot deposits are observed on the inner walls of the quenching probe and are extracted using desiccated gun cleaning patches (Hoppe's 9 1203S). At least two patches are used to collect soot on the inside of the probe. Soot collection on the inside of the probe concludes once an applied patch is observed to have no soot. Similar to PTFE filters, patches are weighed immediately before and 48 h after extraction. The soot mass fraction is determined from the mass of the soot collected from the PTFE filter and gun cleaning patches, m_s, the ratio of the downstream temperature measured at the mass flow controller and time-averaged temperature measurements described in Section 2.2, and the total mass of gas sampled, m_tot.

$$Y_s = \frac{m_\mathrm{s}}{m_\mathrm{tot}} \tag{2}$$

The total mass of gas sampled is 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_gas, the gas sampling time, Δt.

$$m_t = \dot{V} \rho_{gas} \Delta t \frac{T_\infty}{T_\mathrm{g}} \tag{3}$$

Table 1: List of measurements and thermochemical properties of propane burning in a well-ventilated round 37 cm diameter pool fire burning in a quiescent environment.

Fuel	$\dot{m}$	$\dot{Q}$	$\dot{Q}^*$	D^*	L_f	ΔH_c [27]
-	(g/s)	(kW)		(m)	(cm)	(kJ/g)
Propane	0.74 ± 0.01	34.4 ± 0.6	0.37 ± 0.01	0.25 ± 0.01	50.0 ± 16.0	46.34
Propane	0.45 ± 0.01	20.7 ± 0.6	0.22 ± 0.01	0.20 ± 0.01	38.3 ± 14.6	46.34

A description of the soot mass fraction measurement and uncertainty is provided in greater detail in Ref. [8].

2.6 Uncertainty Analysis

An extensive uncertainty analysis of all measurements is provided in Refs. [8, 25, 26]. Measurement uncertainties are estimated using the law of propagation of uncertainty which combines the Type A and B evaluation of uncertainty. For most measurements, the Type A evaluation of uncertainty is the dominant error. Unless otherwise stated, the uncertainty of the measurements are expressed assuming a 95% confidence level.

3. Results and Discussion

3.1 Flame Observations

The measured time-averaged burning rates and calculated heat release rates are provided in Table 1. The heat release rate, $\dot{Q}$, is estimated from the product of the mass burning rate, $\dot{m}$, and the heat of combustion, ΔH_c, of the fuel. The measured mean flame heights, L_f, are observed to match with Heskestad's correlation, shown in Eq. 4, to within the measured uncertainty.

$$\frac{L_\mathrm{f}}{D} = 3.7\,(\dot{Q}^*)^{2/5} - 1.02 \quad ; \quad \dot{Q}^* = \frac{\dot{Q}}{c_p \rho_\infty T_\infty \sqrt{g} D^{5/2}} \tag{4}$$

Here, D is the diameter of the pool fire (30 cm), g is the acceleration of gravity, and c_p and ρ_∞ are the specific heat and the density of air at room temperature, T_∞.

3.2 Comparison of Fire Structure

To compare the propane pool fires of different sizes, the temperature, velocity, and gas species measurements are plotted as a function of the normalized vertical spatial coordinate, z^*:

$$z^* = \frac{z}{D^*} \quad ; \quad D^* = \left(\frac{\dot{Q}}{c_p \rho_\infty T_\infty \sqrt{g}} \right)^{\frac{2}{5}} \tag{5}$$

Here, z is the vertical spatial coordinate, $\dot{Q}$ is the heat release rate, g is the acceleration of gravity, and c_p and ρ_∞ are the specific heat and the density of air at room temperature, T_∞.

Figure 1 displays the time-averaged gas temperature measurements of the 20 kW and 34 kW propane fires as a function of z^*. The time-averaged temperatures of the 20 kW and 34 kW propane fires are observed to peak within the flame ($z^* < 1.32$ [28]) at an approximate z^* of 0.56 and 0.98, respectively. The maximum time-averaged temperature for each pool fire is close to 1325 K.

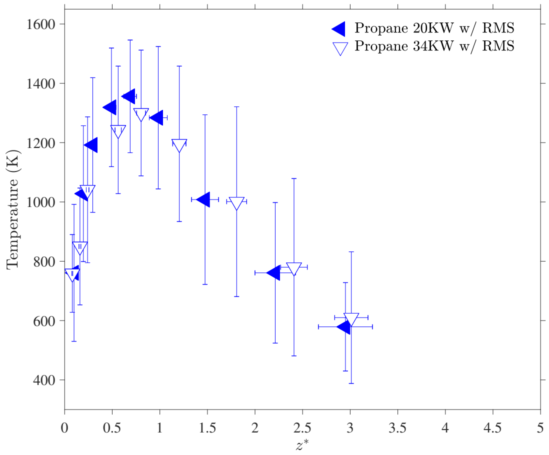


Figure 1: Mean and root mean square (RMS) centerline gas temperature profiles of propane pool fires during their pulsing cycles as a function of z^*.

Time-averaged velocity measurements are presented as a function of z^* in Fig. 2. The velocity of the both propane fires are shown to increase within the flame, then achieve a consistent value (approx. 3.4 m/s) in the intermittent section of the fire ($1.32 < z^* < 3.3$ [28]). When plotted as a function of z^*, the velocity profiles of both propane fires are shown to collapse, suggesting that the flow field within the different size fires evolve in a similar manner.

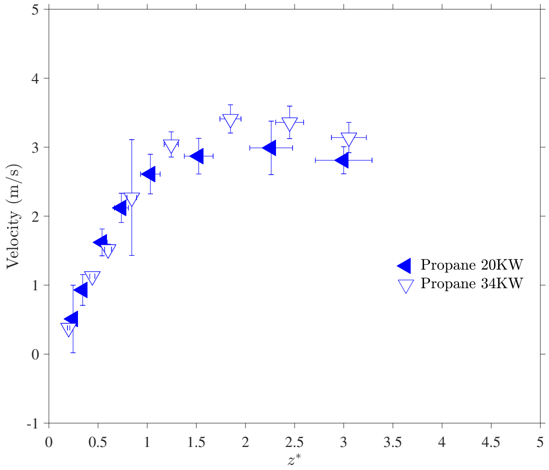


Figure 2: Mean and root mean square (RMS) centerline velocity profiles of and propane pool fires during their pulsing cycles as a function of z^*.

The time-averaged gas volume fraction of major gas species and the mass fraction of soot for the 20 kW and 34 kW propane fires are shown in Figure 3, plotted as a function of z^*. Plots for individual species, including uncertainties, are reported in Ref. [8]. Major species detected in the GC/MSD include combustion reactants (propane, C_3H_8, and oxygen, O_2), combustion products (carbon dioxide, CO_2, and water, H_2O), combustion intermediates (carbon monoxide, CO, and hydrogen, H_2), and inert gases (nitrogen, N_2 and argon, Ar). Trace concentrations of other species are also observed such as methane, ethane, ethylene, acetylene, propene, soot, and benzene.

For both fires, the fuel concentrations are shown to decline as the distance from the fuel surface increases, contrary to oxygen which increases. The intermittent species are observed to peak closer to the fuel surface compared to carbon dioxide and hydrogen. In comparison to the 20 kW propane fire, the 34 kW propane fire is shown to have a higher peak soot mass fraction by approximately a factor of 2.

The carbon to hydrogen ratio is calculated, using Eq. 6, at each measurement location as a way to verify the accuracy of the experimental gas species concentrations measurements.

$$\frac{\mathrm{C}}{\mathrm{H}} = \frac{\sum \mathrm{x}_i \bar{X}_i}{\sum \mathrm{y}_i \bar{X}_i} \tag{6}$$

Here the summation over all measured gas species, and x_i and y_i are the numbers of carbon and hydrogen atoms in the molecule, respectively. For the case of propane, the theoretical carbon to hydrogen ratio is 0.375 and the ratio for each fuel is shown in Fig. 4. As seen in Fig. 4, the theoretical carbon-to-hydrogen ratio for propane, represented by the dotted line, is in agreement with the gaseous species measurements with the experimental uncertainty. The conservation of the carbon to hydrogen ratio at each position for each fuel validates the experimental setup and indicates that the loss of condensable or semi-volatile species during extraction is minimal.

3.3 Carbon Balance

Carbon containing species measured within the propane fires were primarily partitioned into carbon dioxide, carbon monoxide, methane, and soot. Figure 5 displays the mass ratio of carbon monoxide to soot as function of z^* for both propane fires. The general trend of each fuel shows that the ratio of carbon monoxide to soot decreases as the distance from the fuel surface increases. As observed, when plotted as a function of z^*, the mass ratio of carbon monoxide to soot measured from each propane fire of different size tends to collapse, suggesting that carbon is partitioned similarly within propane fires of different size. Köylü et al. [29] reported that the mass-based ratio of CO to soot generation factors in the overfire (i.e., the fuel-lean exhaust stream) region of non-premixed hydrocarbon flames for a range of strongly sooting fuels was 0.34 ± 0.09. It is observed that the ratio of CO to soot tends to Köylü's value at locations higher within fires' plume. Similar results have been previously documented. It should be noted however, that the measurements made here are primarily within the flame envelope, whereas Köylü's are in the overfire region [29].

4. Conclusions

This study characterizes the structure of two medium-scale propane pool fires steadily burning in a quiescent environment. Temperature, velocity, and gas species concentrations for a 20 kW and 34 kW propane fire are reported. The temperature and velocity profiles collapse when plotted as a function of z^*. The calculated carbon-to-hydrogen ratio is shown to be in agreement with the theoretical value at each location. As the fire heat release rate is increased, the time-averaged peak

Sub Topic: Fire Research

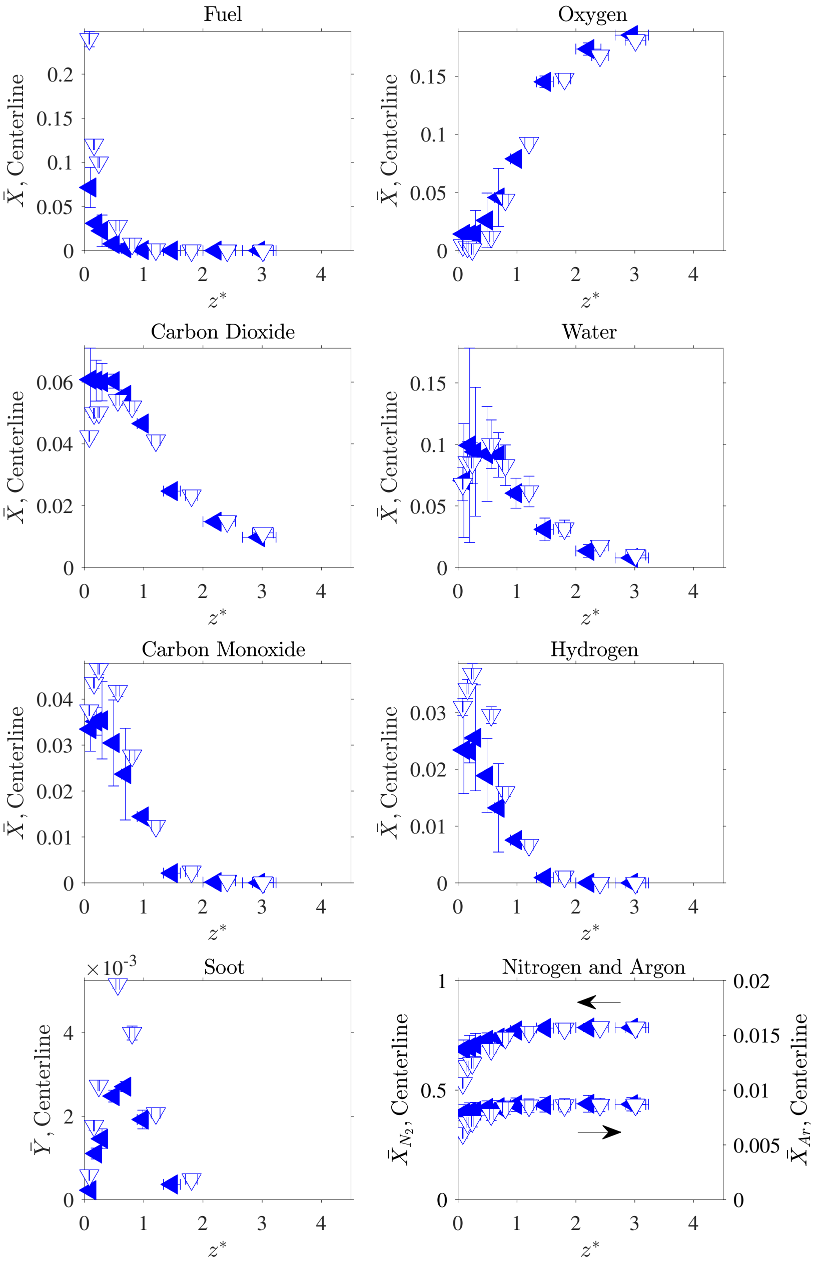


Figure 3: Centerline volume and soot mass fraction profiles of a 20 kW propane (◀) and 34 kW propane (▽) pool fires as a function of z^*. Note, in some cases the presented uncertainty is smaller than the plot markers.

Sub Topic: Fire Research

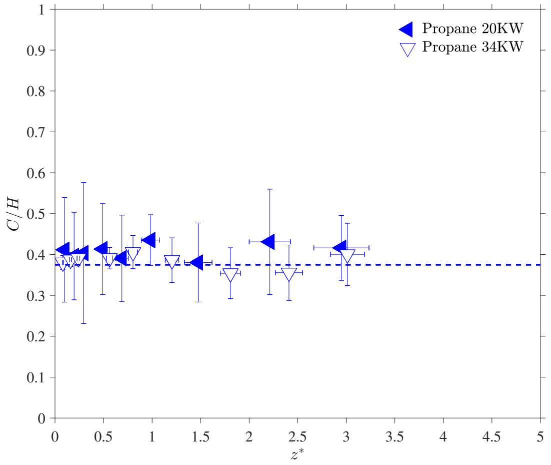


Figure 4: Carbon to hydrogen ratio calculated from all measured gas species compared to the theoretical values as a function of z^*.

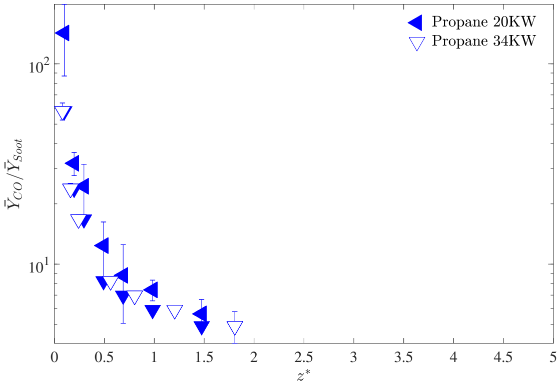


Figure 5: Carbon monoxide to soot ratio as a function of z^*. Note, in some cases the presented uncertainty is smaller than the plot markers.

concentrations on the fire centerline of soot, CO, and H_2 in the fire increase. For a factor of 1.7 increase in the fuel flow, the peak soot centerline concentrations go up by a factor of 2. Robust chemistry sub-models are needed that can address these non-linear finite rate chemistry effects. Future work will apply the same technique to different size propane fires to see if observed trends are consistent. This technique will also be applied to non-centerline positions and pool fires of different fuels to further expand the repository of datasets that will help guide the development and validation of CFD fire models.

Sub Topic: Fire Research

References

[1] S. Fischer, B. Hardouin-Duparc, and W. Grosshandler, The structure and radiation of an ethanol pool fire, Combustion and Flame 70 (1987) 291–306.

[2] A. Hamins and A. Lock, The Structure of a Moderate-Scale Methanol Pool Fire, NIST Technical Note Report No. 1928, National Institute of Standards and Technology, 2016.

[3] M. Choi, G. Mulholland, A. Hamins, and T. Kashiwagi, Comparisons of the soot volume fraction using gravimetric and light extinction techniques, Combustion and Flame 102 (1995) 161–169.

[4] G. Andrews, B. Daham, M. Mmolawa, S. Boulter, J. Mitchell, G. Burrell, J. Ledger, W. Gunamusa, R. Boreham, and H. Phylaktou, FTIR investigations of toxic gases in air starved enclosed fires, Fire Safety Science 8 (2005) 1035–1046.

[5] W. Meier, R. Barlow, Y. Chen, and J. Chen, Raman/Rayleigh/LIF measurements in a turbulent $CH_4/H_2/N_2$ jet diffusion flame: experimental techniques and turbulence–chemistry interaction, Combustion and Flame 123 (2000) 326–343.

[6] M. Bundy, A. Hamins, E. Johnsson, S. Kim, G. Ko, and D. Lenhert, Measurements of heat and combustion products in reduced-scale ventilation-limited compartment fires, NIST Technical Note Report No. 1483, National Institute of Standards and Technology, 2007.

[7] A. Lock, M. Bundy, E. Johnsson, A. Hamins, G. Ko, C. Hwang, P. Fuss, and R. Harris, Experimental Study of the Effects of Fuel Type, Fuel Distribution, and Vent Size on Full-Scale Underventilated Compartment Fires in an ISO 9705 Room, NIST Technical Note Report No. 1603, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2008.

[8] R. Falkenstein-Smith, K. Sung, J. Chen, K. Harris, and A. Hamins, The Structure of Medium-Scale Pool Fires,Version Two, NIST Technical Note Report No. 2082, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2021.

[9] R. Falkenstein-Smith, K. Sung, J. Chen, and A. Hamins, The Chemical Structure of a 30 cm methanol pool fire, Fire and Materials 45 (2021) 429–434.

[10] R. Falkenstein-Smith, K. Sung, J. Chen, and A. Hamins, Chemical Structure of Medium-Scale Liquid Pool Fire, Fire Safety Journal (2020) 103099.

[11] R. Falkenstein-Smith, K. Sung, J. Chen, and A. Hamins, Mixture fraction analysis of combustion products in medium-scale pool fires, Proceedings of the Combustion Institute (2020).

[12] L. Orloff, J. D. Ris, and M. Delichatsios, General correlations of chemical species in turbulent fires, Symposium (International) on Combustion 21 (1988) 101–109.

[13] L. Orloff, J. D. Ris, and M. Delichatsios, Chemical effects on molecular species concentrations in turbulent fires, Combustion and Flame 69 (1987) 273–289.

[14] D. Smith and G. Cox, Major chemical species in buoyant turbulent diffusion flames, Combustion and Flame 91 (1992) 226–238.

[15] R. Falkenstein-Smith, K. Harris, K. Sung, T. Liang, and A. Hamins, A calibration and sampling technique for quantifying the chemical structure in fires using GC/MSD analysis, Fire and Materials (2021).

[16] A. Hamins, S. Fischer, T. Kashiwagi, M. Klassen, and J. Gore, Heat Feedback to the Fuel Surface in Pool Fires, Combustion Science and Technology 97 (1993) 37–62.

[17] 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 86 (1991) 223–228.

[18] A. Hamins, T. Kashiwagi, and R. Buch, Characteristics of pool fire burning, in: Fire Resistance of Industrial Fluids, ASTM International, 1996.

[19] A. Hamins, K. Konishi, P. Borthwick, and T. Kashiwagi, Global properties of gaseous pool fires, Symposium (International) on Combustion 26 (1996) 1429–1436.

[20] C. Shaddix, S. Allendorf, G. Hubbard, D. Ottesen, and L. Gritzo, Diode Laser Diagnostics for Gas Species and Soot in Large Pool Fires, in: SAND2001-8383, 2001.

[21] C. Shaddix, Correcting thermocouple measurements for radiation loss: a critical review, tech. rep. Report No. CONF-990805, Sandia National Laboratories, Livermore, CA, USA, 1999.

[22] F. Incropera, A. Lavine, T. Bergman, and D. DeWitt, Fundamentals of Heat and Mass Transfer, 6th, Wiley, Hoboken, NJ, USA, 2011.

[23] R. Vines, The Platinum Metals and Their Alloys, ed. by E. Wise, Literary Licensing, LLC, 2012.

[24] F. Jaeger and E. Rosenbohm, The exact formulae for the true and mean specific heats of platinum between 0 and 1600 C, Physica 6 (1939) 1123–1125.

[25] K. Sung, J. Chen, M. Bundy, M. Fernandez, and A. Hamins, The thermal character of a 1 m methanol pool fire, NIST Technical Note Report No. 2083, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2020.

[26] K. Sung and A. Hamins, Centerline Velocity Profiles in Medium-Scale Pool Fires, NIST Technical Note Report No. In Progress, National Institute of Standards and Technology, 2021.

[27] M. Hurley, ed., SFPE Handbook of Fire Protection Engineering, 5th, Springer, New York, 2016.

[28] H. R. Baum and B. McCaffrey, Fire induced flow field-theory and experiment, Fire Safety Science 2 (1989) 129–148.

[29] Ü. Ö. Köylü and G. M. Faeth, Carbon monoxide and soot emissions from liquid-fueled buoyant turbulent diffusion flames, Combustion and Flame 87 (1991) 61–76.

Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation

Sarala Padi*
sarala.padi@nist.gov
ITL, NIST
Gaithersburg, MD, USA

Seyed Omid Sadjadi*
omid.sadjadi@nist.gov
ITL, NIST
Gaithersburg, MD, USA

Ram D. Sriram
ram.sriram@nist.gov
ITL, NIST
Gaithersburg, MD, USA

Dinesh Manocha
dmanocha@umd.edu
University of Maryland
College Park, MD, USA

ABSTRACT

Automatic speech emotion recognition (SER) is a challenging task that plays a crucial role in natural human-computer interaction. One of the main challenges in SER is data scarcity, i.e., insufficient amounts of carefully labeled data to build and fully explore complex deep learning models for emotion classification. This paper aims to address this challenge using a transfer learning strategy combined with spectrogram augmentation. Specifically, we propose a transfer learning approach that leverages a pre-trained residual network (ResNet) model including a statistics pooling layer from speaker recognition trained using large amounts of speaker-labeled data. The statistics pooling layer enables the model to efficiently process variable-length input, thereby eliminating the need for sequence truncation which is commonly used in SER systems. In addition, we adopt a spectrogram augmentation technique to generate additional training data samples by applying random time-frequency masks to log-mel spectrograms to mitigate overfitting and improve the generalization of emotion recognition models. We evaluate the effectiveness of our proposed approach on the interactive emotional dyadic motion capture (IEMOCAP) dataset. Experimental results indicate that the transfer learning and spectrogram augmentation approaches improve the SER performance, and when combined achieve state-of-the-art results.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Machine learning**;

KEYWORDS

Attentive pooling, IEMOCAP, ResNet, spectrogram augmentation, speech emotion recognition (SER), transfer learning

*Both authors contributed equally to this research.

ACM Reference Format:
Sarala Padi, Seyed Omid Sadjadi, Ram D. Sriram, and Dinesh Manocha. 2021. Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3462244.3481003

1 INTRODUCTION

Automatic emotion recognition plays a key role in human-computer interaction where it can enrich the next-generation AI with emotional intelligence by grasping the emotion from voice and words [32, 37]. The motivation behind developing algorithms to analyze emotions is to design computer interfaces that mimic and embed realistic emotions in synthetically generated responses [6]. Furthermore, research studies have shown that emotions play a critical role in the decision-making process for humans [6]. Hence, there is a growing demand to develop automatic systems that understand and recognize human emotions.

Humans express emotions in several ways, and speech is considered the most effective communication method to express feelings. For speech emotion recognition (SER), traditionally, machine learning (ML) models were developed using hand-crafted and engineered features such as mel-frequency cepstral coefficients (MFCC), Chroma-based features, pitch, energy, entropy, and zero-crossing rate [16, 21, 43], to mention a few. However, the performance of such ML models depends on the type and diversity of the features used. Although it remains unclear which features correlate most with various emotions, the research is still ongoing to explore additional features and new algorithms to model the dynamics of feature streams representing human emotions. On the other hand, the recent advancements in deep learning, along with the available computational capabilities, have enabled the research community to build end-to-end systems for SER. A big advantage of such systems is that they can directly learn the features from spectrograms or raw waveforms [12, 23, 36, 41, 45], thereby obviating the need for extracting a large set of hand-crafted features [13]. Recent studies have proposed the use of convolutional neural network (CNN) models combined with long short-term memory (LSTM) built on spectrograms and raw waveforms, showing improved SER performance [19, 23, 24, 26, 35, 36, 46]. However, building such complex systems requires large amounts of labeled training data. Also, the insufficient labeled training data can potentially make the models

Padi, Sarala; Sadjadi, Omid; Sriram, Ram D.; Manocha, Dinesh. "Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation." Presented at ACM International Conference on Multimodal Interaction, Montreal, CA. October 18, 2021 - October 22, 2021.

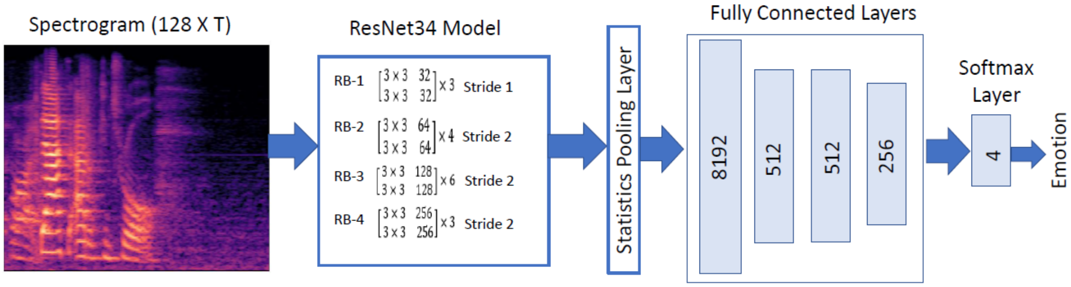

Spectrogram (128 X T)
ResNet34 Model
RB-1 $\begin{bmatrix}3\times3 & 32\\3\times3 & 32\end{bmatrix}\times3$ Stride 1
RB-2 $\begin{bmatrix}3\times3 & 64\\3\times3 & 64\end{bmatrix}\times4$ Stride 2
RB-3 $\begin{bmatrix}3\times3 & 128\\3\times3 & 128\end{bmatrix}\times6$ Stride 2
RB-4 $\begin{bmatrix}3\times3 & 256\\3\times3 & 256\end{bmatrix}\times3$ Stride 2
Statistics Pooling Layer
Fully Connected Layers
8192
512
512
256
Softmax Layer
4
Emotion

Figure 1: Block diagram of the proposed SER system. T denotes the number of frames.

overfit to specific data conditions and domains, resulting in poor generalization on unseen data.

This paper addresses the insufficient data problem using a transfer learning approach combined with a spectrogram augmentation strategy. We re-purpose a residual network (ResNet) model [17] developed for speaker recognition using large amounts of speaker-labeled data and use it as a feature descriptor for SER. The model includes a statistics pooling layer that enables processing of variable length segments without a need for truncation. Also, we increase the training data size by generating more data samples using spectrogram augmentation [30]. We evaluate the effectiveness of our proposed system on the interactive emotional dyadic motion capture (IEMOCAP) dataset [3].

2 RELATED WORK

Recently, neural network based modeling approaches along with different variations of attention mechanism (e.g., plain [18], local [24], and self [40]) have shown promise for SER. Among them, techniques such as bidirectional LSTMs (BLSTM) [10, 18, 24, 31, 42] and time-delay neural networks (TDNN) [44], which can effectively model relatively long contexts compared to their DNN counterparts, have been successfully applied for SER on the IEMOCAP. Nevertheless, as discussed previously, the lack of large amounts of carefully labeled data to build complex models for emotion classification remains a main challenge in SER [1]. To address this, two approaches are commonly used: data augmentation and transfer learning.

Data augmentation methods generate additional training data by perturbing, corrupting, mimicking, and masking the original data samples to enable the development of complex ML models. For example, [4, 29, 35] applied signal-based transformations such as speed perturbation, time-stretch, pitch shift, as well as added noise to original speech waveforms. One disadvantage of these approaches is that they require signal-level modifications, thereby increasing the computational complexity and storage requirements of the subsequent front-end processing. They can also lead to model overfitting due to potentially similar samples in the training set, while random balance can potentially remove useful information [4]. For example, in [9] a vocal tract length perturbation (VTLP) approach was explored for data augmentation along with a CNN model, and was found to result in a lower accuracy compared to a baseline model due to overfitting issues.

Since generative adversarial network (GAN) based models have demonstrated remarkable success in computer vision, several studies have recently incorporated this idea to address the data scarcity problem and to generate additional data samples for SER [4, 8]. For instance, [4] addressed the data imbalance using signal-based transformations and GAN based models for generating high-resolution spectrograms to train a VGG19 model for an emotion classification task and showed that GAN-generated spectrograms outperformed signal-based transformations. However, GAN generated features are strongly dependent on the data used during training and may not generalize to other datasets. Another challenge with GAN-based augmentation is that it is difficult to train and optimize.

Another effective way to address challenges related to data scarcity is transfer learning [2, 11, 14, 28, 49]. Transfer learning can leverage the information and knowledge learned from one related task and domain to another. Several recent studies have proposed transfer learning methods to improve SER performance and have shown these methods to outperform prior methods in recognizing emotions even for unseen scenarios, individuals, and conditions [15]. It has been shown that transfer learning can increase feature learning abilities, and that the transferred knowledge can further enhance the SER classification accuracy [7, 13, 22, 39]. To further improve the SER performance, transfer learned features have been used in combination with deep belief networks (DBN) [22], recurrent neural networks (RNN) [13], CNN [39], temporal convolutional network (TCN) [49], and sparse autoencoder [7]. However, transfer learning methods have not been fully explored and analyzed for emotion recognition. Particularly, it is unclear whether and how ML models trained for other data-rich speech applications such as speaker recognition would perform for SER.

3 PROPOSED SYSTEM

Figure 4 shows the block diagram of the proposed system for speech emotion recognition. We use an end-to-end system with a ResNet34 model [17] to perform emotion classification. ResNet models, originally developed for computer vision applications [17], have recently gained interest for speech applications such as speaker recognition [48]. The residual blocks introduced in ResNet models allow us to train much deeper models that are otherwise difficult, if not impossible, to train due to vanishing and exploding gradient problems. The ResNet models also allow the higher layers to learn the identity function so that higher-level features perform equally well on

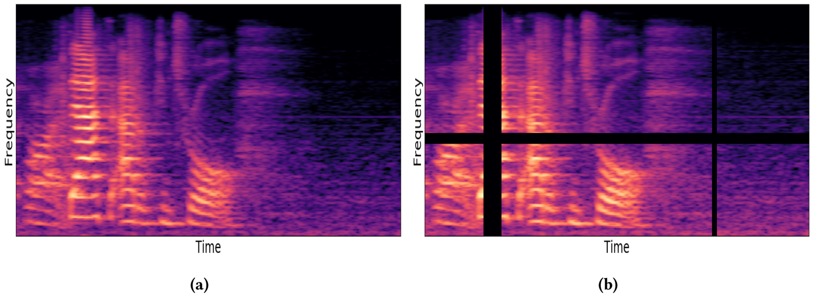


Figure 2: (a) original spectrogram, and (b) spectrogram modified using (multiple) masking blocks of consecutive time steps (vertical masks) and mel frequency channels (horizontal masks). The black horizontal and vertical stripes indicate the masked portions of the spectrogram.

unseen data compared to the lower layers of the model. In our proposed system, the convolutional layers in the model learn feature representations (feature maps) and reduce the spectral variations into compact representations, while the fully connected (FC) layers take the contextual features and generate predictions for emotion classification.

3.1 Input data

Although SER systems traditionally used a large set of low-level time- and frequency-domain features to capture and represent the various emotions in speech, in recent years many state-of-the-art SER systems use complex neural network models that learn directly from spectrograms, or even raw waveforms. Accordingly, in this study, we build and explore a ResNet based system using log-mel spectrograms as input features. We extract high-resolution spectrograms to enable the model to not only learn the spectral envelope structure, but also the coarse harmonic structure for the various emotions.

3.2 Transfer learning

As noted previously, transfer learning is a ML method where a model initially developed for one task or domain is re-purposed, partly or entirely, for a different but related task/domain. It has recently gained interest for SER [11]. In this study, we re-purpose a model initially developed for speaker recognition to serve as a feature descriptor for SER. More specifically, we first train a ResNet34 model on large amounts of speaker-labeled audio data. Then, we replace the FC layers of the pre-trained model with new randomly initialized FC layers. Finally, we re-train the new FC layers for an SER task on the IEMOCAP dataset.

3.3 Statistics pooling

As shown in Figure 4, the proposed system employs a statistics pooling layer [38] that aggregates the frame-level information over time and reduces the sequence of frames to a single vector by concatenating the mean and standard deviation computed over frames. Accordingly, the convolutional layers in the ResNet model work at the frame-level, while the FC layers work at the segment-level. This enables the system to efficiently model variable-length sequences of frames, thereby eliminating the need for truncating the sequence of frames to a pre-specified length to match that of the segments used during training. The sequence-truncation approach, which is commonly adopted in neural network based SER systems, can have a deleterious impact on SER performance as potentially informative frames are dropped out from the input. It is worth noting here that the statistics pooling can be viewed as an attention mechanism with equal weights for all frames, which also appends second order statistics (i.e., standard deviation) to capture long-term temporal variability over the duration of segments.

Table 1: Parameter settings for the conservative and aggressive augmentation policies. Here, N_f and N_t denote the number of frequency and time masks applied.

Augmentation Policy	F	W	p	N_f	N_t
None	0	0	–	–	–
Conservative	15	50	0.2	2	2
Aggressive	27	70	0.2	2	2

3.4 Spectrogram augmentation

Currently, the majority of the features and methods for SER are adapted from speech recognition, speaker recognition, or speech synthesis fields [20]. There has been recent success in applying a computationally efficient data augmentation strategy, termed spectrogram augmentation, for speech recognition tasks [30]. The spectrogram augmentation technique generates additional training data samples by applying random time-frequency masks to spectrograms to mitigate the overfitting issue and improve the generalization of speech recognition models. Motivated by promising results seen with the spectrogram augmentation in the speech recognition field, we augment the training data using spectro-temporally modified versions of the original spectrograms (see Figure 2). Because the time-frequency masks are applied directly to spectrograms, the augmentation can be conveniently applied on-the-fly, eliminating the necessity to create and store new data files as commonly done in many augmentation approaches for speech applications.

Similar to the approach taken in [30], we consider two policies to systematically apply spectrogram augmentation for SER, namely conservative and aggressive. The frequency masking is applied over f consecutive frequency channels in the range $[f_0, f_0 + f)$, where f is sampled from a uniform distribution $[0, F]$ and f_0 is sampled from $[0, \nu - f]$. Here, F and ν denote the maximum width of frequency masks and the total number of frequency channels, respectively. The time masking, on the other hand, is applied over t consecutive frames in the range $[t_0, t_0 + t)$, where t is selected from a uniform distribution $[0, W]$ and t_0 is sampled from $[0, T - t]$. Similarly, W and T denote the maximum width of time masks and the number of time frames, respectively. An upper bound is also applied on the width of the time masks such that $W = \min(W, pT)$, i.e., the width of a mask cannot be longer than p times the number of time frames. This is to ensure sufficient speech content after masking, in particular for shorter segments. Table 1 summarizes the various parameters for the two spectrogram augmentation policies used in this paper.

4 EXPERIMENTS

4.1 Dataset

We evaluate the effectiveness of the proposed SER system on the IEMOCAP dataset [3], which contains improvised and scripted multimodal dyadic conversations between actors of opposite gender. It consists of 12 hours of speech data from 10 subjects, pre-segmented into short cuts that were judged by three annotators to generate emotion labels. It includes nine categorical emotions and 3-dimensional labels. In our experiments, we only consider the speech segments for which at least two annotators agree on the emotion label. In an attempt to replicate the experimental protocols used in a number of prior studies, we conduct three experiments on the full dataset (i.e., the combined improvised and scripted portions): Exp 1, using four categorical emotions: "angry", "happy", "neutral", "sad"; Exp 2, using the same categories as in Exp 1, but replacing the "happy" category with "excited"; Exp 3, by merging the "happy" and "excited" categories from Exp 1 and Exp 2. The total number of examples used for Exp 1 is 4490 and the number of examples per category is 1103, 595, 1708, and 1084, respectively. The number of examples in the "excited" category is 1041, making the total number of examples in the merged category (i.e., Exp 3) 1636. Table 2 summarizes the data statistics in the IEMOCAP dataset for the three experimental setups considered in this study.

The IEMOCAP dataset comprises five sessions, and the speakers in the sessions are non-overlapping. Therefore, there are 10 speakers in the dataset, i.e., 5 female and 5 male speakers. To conduct the experiments in a speaker-independent fashion, we use a leave-one-session-out (LOSO) cross-validation strategy, which results in 5 different train-test splits/folds. For each fold, we use the data from 4 sessions for training and the remaining one session for model evaluation. Since the dataset is multi-label and imbalanced, in addition to the overall accuracy, termed weighted accuracy (WA), we report the average recall over the different emotion categories, termed unweighted accuracy (UA), to present our findings. Additionally, to understand and visualize the performance of the proposed system within and across the various emotion categories, we compute and report confusion matrices for the three experiments. Note that for each experiment, we compute the average of performance metrics over the five training-test splits as the final result.

Table 2: Data statistics for the various emotion classes in the IEMOCAP for the three experimental setups considered in this study. Both the improvised and scripted portions of the IEMOCAP dataset are used in our experiments.

Experiment	Emotion	#segments
Exp 1	Angry	1103
	Happy	595
	Neutral	1708
	Sad	1084
	Total	**4490**
Exp 2	Angry	1103
	Excited	1041
	Neutral	1708
	Sad	1084
	Total	**4936**
Exp 3	Angry	1103
	Excited+Happy	1636
	Neutral	1708
	Sad	1084
	Total	**5531**

4.2 Setup and configuration

For speech parameterization, high resolution 128-dimensional log-mel spectrograms are extracted from 25 ms frames at a 100 Hz frame rate (i.e., every 10 ms). For feature normalization, a segment level mean and variance normalization is applied[1]. Note that this is not ideal as typically the normalization is applied at the recording/conversation level. We have found that normalizing the segments using statistics computed at the conversation level significantly improves the SER performance on the IEMOCAP. Nevertheless, this violates the independence assumption for the speech segments, hence it is not considered in this study. The front-end processing, including feature extraction and feature normalization, is performed using the NIST speaker and language recognition evaluation (SLRE) [33, 34] toolkit. While training the model, we select T-frame chunks using random offsets over original speech segments where T is randomly sampled from the set $\{150, 200, 250, 300\}$ for each batch. For speech segments shorter than T frames, signal padding is applied. On the other hand, while evaluating the model, we feed the entire duration of the test segments because the statistics pooling layer enables the model to consume variable-length inputs.

As noted previously, the proposed end-to-end SER system uses a pre-trained ResNet34 model built on a speaker recognition task. We train the ResNet34 model on millions of speech samples from more than 7000 speakers available in the VoxCeleb corpus [25]. To build the speaker recognition model, we apply the same front-end processing described above to extract high-resolution log-mel

[1] No voice activity detection (VAD) is applied prior to feature normalization because it was found to be detrimental to SER performance on the IEMOCAP. We hypothesize that this is because the silence gaps in within and between utterances might be relevant in terms of speakers' emotional state.

Table 3: Performance comparison of our proposed approach with prior methods that use the LOSO strategy for experiments on the full IEMOCAP dataset (i.e., both the improvised and scripted portions). Abbreviations: A-Angry, H- Happy, N-Neutral, E-Excited, H+E: Happy and Excited merged. Blanks (–) indicate unreported values.

Experiment (emotion classes)	Approach	UA [%]	WA [%]
Exp 1 (A, H, S, N)	BLSTM+attention [18]	49.96	59.33
	Transformer+self-attention [40]	58.01	59.43
	BLSTM+local attention [24]	58.8	63.5
	BLSTM+attention [31]	59.6	62.5
	Proposed	**61.61**	**66.02**
Exp 2 (A, E, S, N)	CTC-BLSTM [5]	54	–
	BLSTM+attention [42]	55.65	–
	Transformer+self-attention [40]	64.79	64.33
	Proposed	**65.56**	**65.62**
Exp 3 (A, H+E, S, N)	BLSTM+transfer learnig [13]	51.86	50.47
	VGG19+GAN augmentation [4]	54.6	–
	CNN+attention+multi-task learning [26]	–	56.10
	BLSTM+self-attention [10]	57.0	55.7
	CNN+attention+multi-task/transfer learning [27]	59.54	–
	ResTDNN+self-attention [44]	61.32	60.64
	Proposed	**64.14**	**63.61**

spectrograms from VoxCeleb data. We conduct experiments using models with and without transfer learning and spectrogram augmentation. For each original speech segment, we generate and augment two spectro-temporally modified versions according to the augmentation policies defined in Table 1. This is applied for both speaker and emotion recognition systems during training. To study the impact of the statistics pooling layer, we also evaluate these models with and without this layer. For all the experiments, we use a categorical cross-entropy loss as the objective function to train the models. The number of channels in the first block of the ResNet model is set to 32. The model is trained using Pytorch[2] and the stochastic gradient descent (SGD) optimizer with momentum (0.9), an initial learning rate of 10^{-2}, and a batch size of 32. The learning rate remains constant for the first 8 epochs, after which it is halved every other epoch. We use parametric rectified linear unit (PReLU) activation functions in all layers (except for the output), and utilize a layer-wise batch normalization to accelerate the training process and improve the generalization properties of the model.

5 RESULTS

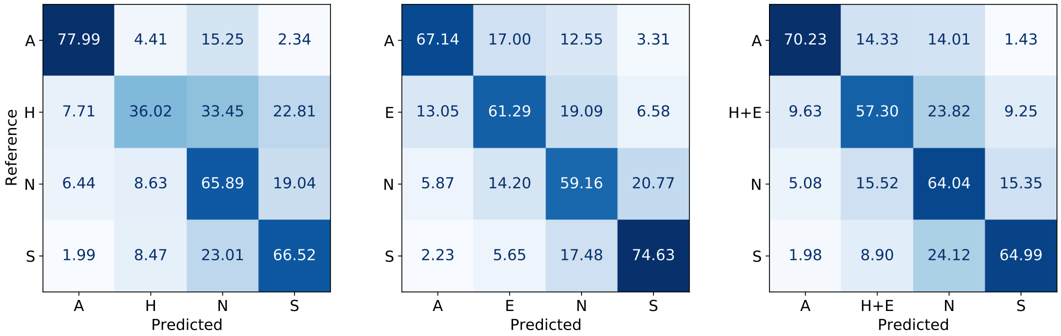
Table 3 presents the performance comparison of our proposed system with several prior approaches for the three experimental setups (i.e., Exp 1, 2, and 3) described in Section 4. The results are obtained using the combined system that utilizes the ResNet model with the statistics pooling layer trained using the transfer learning and spectrogram augmentation approaches described in Section 3. All studies referenced in the table adopt the LOSO strategy to conduct experiments on both the improvised and scripted portions of the IEMOCAP dataset[3]. It can be seen from the table that the proposed system consistently provides competitive performance across the three experiments, achieving state-of-the-art results. In the case of Exp 2, the proposed system outperforms a system that uses 384 engineered features [40], while for the other two experiments, our proposed system outperforms systems that use a large set of engineered features (e.g., [24] and [31]).

To visualize the performance of the proposed system within and across the different emotion categories, confusion matrices for the three experimental setups are shown in Figure 3. It is observed from Figure 3(a) that the system confuses the "happy" class (H) with the "neutral" class (N) quite often, while performing the best on the "angry" class (A). This is consistent with observations reported in other studies on IEMOCAP [26, 47]. Our informal listening experiments confirm that the "happy" and "neutral" classes are indeed confusable emotion pairs in the IEMOCAP dataset. The system performance balance is improved in Figure 3(b) where we replace the less pronounced "happy" category with the "excited" category (E). Combining the "happy" and "excited" categories in Exp 3 further improves the performance balance across the various emotions, at the expense of increasing the confusion between the "angry" (A) and "excited" plus "happy" (H+E) categories.

To investigate and quantify the contribution of the various system components proposed in this study for improved SER, we further conduct ablation experiments to measure the system performance with and without the transfer learning, the spectrogram augmentation, and the statistics pooling layer. For these ablation

[2] https://github.com/pytorch/pytorch

[3] There are other related studies in the literature that only use the improvised portion of the IEMOCAP dataset [9, 36, 49]. On the other hand, in our experiments, we use both the improvised and scripted portions of the IEMOCAP, which is approximately twice the size of the improvised portion alone. Because the experimental setups and the amount of data used for model training and evaluation in those studies are different than ours, we have not included them in Table 3 for comparison. The SER performance on the improvised portion is known to be better than that on the full dataset (e.g., see [13, 26, 31, 40, 42]).



(a) Exp 1: A, H, N, S

(b) Exp 2: A, E, N, S

(c) Exp 3: A, H+E, N, S

Figure 3: Performance confusion matrices of the proposed SER system for the three experiments conducted in this study using the LOSO strategy. Abbreviations: A-angry, E-excited, H-happy, N-neutral, S-Sad, and H+E- Happy and Excited merged.

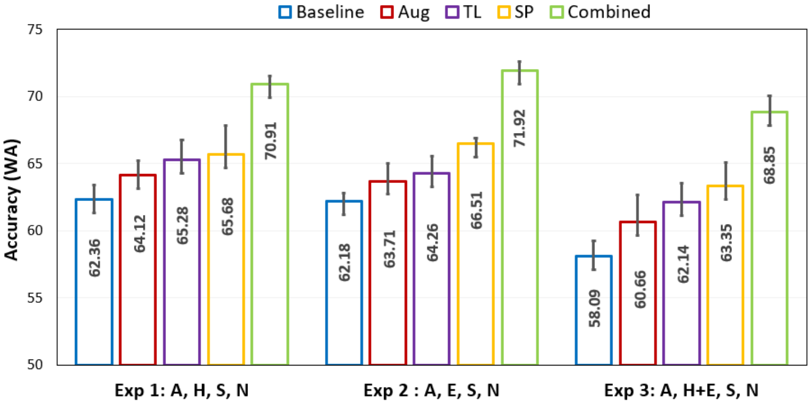


Figure 4: Performance (WA) of the proposed approach with and without transfer learning (TL), spectrogram augmentation (Aug), and statistics pooling (SP). All results are obtained using a 5-fold cross-validation. The height of each bar represents the average accuracy computed over 5 runs, while the error bars denote the standard deviations over the 5 runs. Abbreviations: A-Angry, H-Happy, N-Neutral, S-Sad, E-Excited, H+E: Happy and Excited merged

experiments, we employ a 5-fold cross-validation (CV) strategy, where we use 80% of the data for training and 20% for testing the system. This process is repeated 5 times to reduce possible partition-dependencies. Figure 4 shows the average overall classification accuracy (WA) computed across 5 folds (or 5 runs). The height of the bars represents the average accuracy, and the error bars denote the standard deviations computed over the 5 runs. It is observed that the proposed components, both individually and in combination, consistently provide performance gains across the three experimental setups (i.e., Exp 1, 2 and 3). The statistics pooling approach seems to have the greatest impact on performance, followed by the transfer learning and spectrogram augmentation methods. Furthermore, the model that combines all the system components not only consistently achieves the best performance, but also relatively smaller variation across the 5 runs as evidenced by the error bars.

6 CONCLUSIONS

In this paper, we explored a transfer learning approach along with a spectrogram augmentation strategy to improve the SER performance. Specifically, we re-purposed a pre-trained ResNet model from speaker recognition that was trained using large amounts of speaker-labeled data. The convolutional layers of the ResNet model were used to extract features from high-resolution log-mel spectrograms. In addition, we adopted a spectrogram augmentation technique to generate additional training data samples by applying random time-frequency masks to log-mel spectrograms to mitigate overfitting and improve the generalization of emotion recognition models. We evaluated the proposed system using three different experimental settings and compared the performance against that of several prior studies. The proposed system consistently provided competitive performance across the three experimental setups, achieving state-of-the-art results on two settings. The state-of-the-art results were achieved without the use of engineered features. It was also shown that incorporating the statistics pooling layer to accommodate variable-length audio segments improved the emotion recognition performance. Results from this study suggest that, for practical applications, simplified front-ends with only spectrograms can be as effective for SER, and that models trained for data-rich speech applications such as speaker recognition can be re-purposed using transfer learning to improve the SER performance under data scarcity constraints. In the future, to further enhance the emotion recognition accuracy, we will extend our work along these lines by exploring more data augmentation methods, incorporating other transfer learning paradigms, and evaluating the proposed system across different datasets.

7 ACKNOWLEDGEMENT

Experiments and analyses were performed, in part, on the NIST Enki HPC cluster.

8 DISCLAIMER

The views and conclusions presented in this paper are those of the authors and should not be interpreted as the official findings, either expressed or implied, of NIST or the U.S. Government.

REFERENCES

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *Proc. ACM ICM.* 292–301.

[2] George Boateng and Tobias Kowatsch. 2020. Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning. In *Proc. ACM ICMI.* 12–16.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[4] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. 2019. Data augmentation using GANs for speech emotion recognition. In *Proc. INTERSPEECH.* 171–175.

[5] Vladimir Chernykh and Pavel Prikhodko. 2017. Emotion recognition from speech with recurrent neural networks. *arXiv preprint arXiv:1701.08071* (2017).

[6] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18, 1 (2001), 32–80.

[7] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. 2013. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction.* 511–516.

[8] Sefik Emre Eskimez, Dimitrios Dimitriadis, Robert Gmyr, and Kenichi Kumanati. 2020. GAN-based data generation for speech emotion recognition. *Proc. INTERSPEECH* (2020), 3446–3450.

[9] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. 2018. CNN+LSTM architecture for speech emotion recognition with data augmentation. *arXiv preprint arXiv:1802.05630* (2018).

[10] Han Feng Feng, Sei Uno, and Tatsuya Kawahara. 2020. End-to-End speech emotion recognition combined with acoustic-to-word ASR model. In *Proc. INTERSPEECH.* 501–505.

[11] Kexin Feng and Theodora Chaspari. 2020. A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science* 2, 9 (2020).

[12] Mengna Gao, Jing Dong, Dongsheng Zhou, Qiang Zhang, and Deyun Yang. 2019. End-to-end speech emotion recognition based on one-dimensional convolutional neural network. In *Proc. ACM ICIAI.* 78–82.

[13] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation Learning for Speech Emotion Recognition. In *Proc. INTERSPEECH.* 3603–3607.

[14] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Progressive neural networks for transfer learning in emotion recognition. In *Proc. INTERSPEECH.*

[15] John Gideon, Melvin McInnis, and Emily Mower Provost. 2019. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Trans. Affective Computing* (2019).

[16] Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Proc. INTERSPEECH.* 223–227.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[18] Che-Wei Huang and Shri Narayanan. 2016. Attention assisted discovery of sub-utterance structure in speech emotion recognition. In *Proc. INTERSPEECH.* 1387–1391.

[19] Gil Keren and Björn Schuller. 2016. Convolutional RNN: an enhanced model for extracting features from sequential data. In *Proc. IEEE IJCNN.* 3412–3419.

[20] Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International Journal of Speech Technology* 15, 2 (2012), 99–117.

[21] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. Emotion recognition by speech signals. In *Proc. INTERSPEECH.* 125–128.

[22] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. 2018. Transfer learning for improving speech emotion classification accuracy. In *Proc. INTERSPEECH.* 257–261.

[23] Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. 2018. Emotion recognition from variable-length speech segments using deep learning on spectrograms. In *Proc. INTERSPEECH.* 3683–3687.

[24] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proc. IEEE ICASSP.* 2227–2231.

[25] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 1–15.

[26] Michael Neumann and Ngoc Thang Vu. 2017. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In *Proc. INTERSPEECH.* 1263–1267.

[27] Michael Neumann and Ngoc Thang Vu. 2019. Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech. In *Proc. IEEE ICASSP.* 7390–7394.

[28] Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller. 2020. Group-level speech emotion recognition utilising deep spectrum features. In *Proc. ACM ICMI.* 821–826.

[29] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak. 2020. X-Vectors meet emotions: A study on dependencies between emotion and speaker recognition. In *Proc. IEEE ICASSP.* 7169–7173.

[30] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proc. INTERSPEECH.* 2613–2617.

[31] Gaetan Ramet, Philip N Garner, Michael Baeriswyl, and Alexandros Lazaridis. 2018. Context-aware attention mechanism for speech emotion recognition. In *Proc. IEEE SLT Workshop.* 126–131.

[32] Harper Richard, R Tom, R Yvonne, and S Abigail. 2008. *Being Human: Human-Computer Interaction in The Year 2020.* Report, Microsoft Corporation.

[33] Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, and Jaime Hernandez-Cordero. 2020. The 2019 NIST audio-visual speaker recognition evaluation. In *Proc. Speaker Odyssey Workshop.* 259–265.

[34] Seyed Omid Sadjadi, Timothee Kheyrkhah, Audrey Tong, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, and Jaime Hernandez-Cordero. 2018. The 2017 NIST language recognition evaluation. In *Proc. Speaker Odyssey Workshop.* 82–89.

[35] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak. 2018. Emotion identification from raw speech signals using DNNs. In *Proc. INTERSPEECH.* 3097–3101.

[36] Aharon Satt, Shai Rozenberg, and Ron Hoory. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *Proc. INTERSPEECH.* 1089–1093.

[37] Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 5 (2018), 90–99.

[38] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. IEEE ICASSP.* 5329–5333.

[39] Peng Song, Yun Jin, Li Zhao, and Minghai Xin. 2014. Speech emotion recognition using transfer learning. *IEICE Trans. Information and Systems* 97, 9 (2014), 2530–2532.

[40] Lorenzo Tarantino, Philip N Garner, and Alexandros Lazaridis. 2019. Self-attention for speech emotion recognition. In *Proc. INTERSPEECH*. 2578–2582.

[41] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. IEEE ICASSP*. 5200–5204.

[42] Samarth Tripathi, Tripathi Sarthak, and Homayoon Beigi. 2018. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *arXiv preprint arXiv:1804.05788* (2018).

[43] Dimitrios Ververidis and Constantine Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication* 48, 9 (2006), 1162–1181.

[44] Wen Wu, Chao Zhang, and Philip C. Woodland. 2021. Emotion recognition by fusing time synchronous and time asynchronous representations. In *Proc. IEEE ICASSP*. 6269–6273.

[45] Zixiaofan Yang and Julia Hirschberg. 2018. Predicting arousal and valence from waveforms and spectrograms using deep neural networks. In *Proc. INTERSPEECH*. 3092–3096.

[46] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. 2018. Speech emotion recognition using spectrogram & phoneme embedding. In *Proc. INTERSPEECH*. 3688–3692.

[47] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *Proc. IEEE SLT Workshop*. IEEE, 112–118.

[48] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. 2019. BUT system description to VoxCeleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592* (2019).

[49] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Shihuang Sun, Haishuai Wang, Jianhua Tao, and Björn W. Schuller. 2021. Self-attention transfer networks for speech emotion recognition. *Virtual Reality & Intelligent Hardware* 3, 1 (2021), 43–54.

Sub topic: Fire Research

12th U. S. National Combustion Meeting
Organized by the Central States Section of the Combustion Institute
May 24–26, 2021 (Virtual)
College Station, Texas

Investigating Firebrand Deposition Processes in Large Outdoor Fires

Sayaka Suzuki[1] and Samuel L. Manzello[2]

[1]National Research Institute of Fire and Disaster (NRIFD), JAPAN
[2]National Institute of Standards and Technology (NIST), USA

Corresponding author: sayakas@fri.go.jp

Abstract: Devastating large outdoor fires have been responsible for destruction of vast amounts of infrastructure and loss of human life. Wildland fires that spread into urban areas, known as wildland-urban interface (WUI) fires, are capable of enormous destruction. WUI fires are distinct from wildland fires; WUI fires include the combustion of both vegetative and human-made fuels and occur where large population centers exist whereas wildland fires include the combustion of vegetative fuels and occur in uninhabited areas. The rise of densely populated urban areas has also seen the development of large urban fires. The most recent of these occurred in the winter of 2016 in Niigata, Japan. Similarly, the USA has also experienced several major urban fires. In some cases, earthquakes have served to initiate these fires, but it is not a necessary condition for these urban fires to develop. In addition, the rise of informal settlement communities in Southeast Asia and Africa continues to result in large outdoor fires capable of great destruction. Firebrands, or smoldering and/or flaming particles, are in fact the main culprit to destroy structures in large outdoor fires. Recent comprehensive review of firebrand combustion reported that deposition processes remain largely unexplored. As part of this work, a series of intricate experiments were undertaken to investigate firebrand deposition processes. A firebrand generator was utilized, and various flow obstructions were placed down stream of these firebrand generators to better understand these complex deposition processes. Results of these investigations for multiple wind speeds, firebrand size and mass distributions, and obstacle placement are presented and discussed.

Keywords: large outdoor fires; firebrands; wind; deposition

1. INTRODUCTION

Wildfires that spread into communities, referred to as Wildland-Urban Interface (WUI) fires, are a dilemma throughout the world [**1**]. Japan has experienced large urban fires. Some after strong earthquakes, such as 1997 Hanshin-Awaji Earthquake. Yet, it is not a necessary condition for these urban fires to develop.

A major factor in both WUI and urban fire spread, is firebrand production [**2**]. When structures burn in these fires, pieces of burning material, known as firebrands, are generated, become lofted, and are carried by the wind. This results in showers of wind-driven firebrands. These spot fires overwhelm firefighting resources [**1-2**].

In Japan, several city fire spread models were developed for damage estimation. These models were based on the past city fire damages and use empirical formula under limited situations

[**3**]. There is a lack of scientific and reliable data needed to advance such models further; especially for firebrand accumulation.

As structures are exposed to wind, stagnation planes are produced around structures. The authors demonstrated that firebrands may accumulate in these stagnation planes [**4**]. In a subsequent study performed by the authors [**5**], a series of full-scale experiments revealed that wind speed influences not only the spatial location and extent of the accumulated firebrands in the stagnation plane in front of the obstacle, but also the nature of the smoldering combustion intensity of the accumulated firebrands. This paper describes an in-depth study of this phenomenon at reduced-scale to determine if useful insights may be obtained from simpler experiments. The firebrand distributions were varied to simulate *both* burning structures and vegetation, as prior studies were focused on vegetative firebrands [**4-5**]. The paper closes with a brief comparison to full-scale experiments described in Suzuki and Manzello [**5**].

2. EXPERIMENTAL DESCRIPTION

A reduced-scale continuous firebrand generator was used to generate firebrand showers (**Figure 1**). The following experimental description follows those presented elsewhere [**6**].

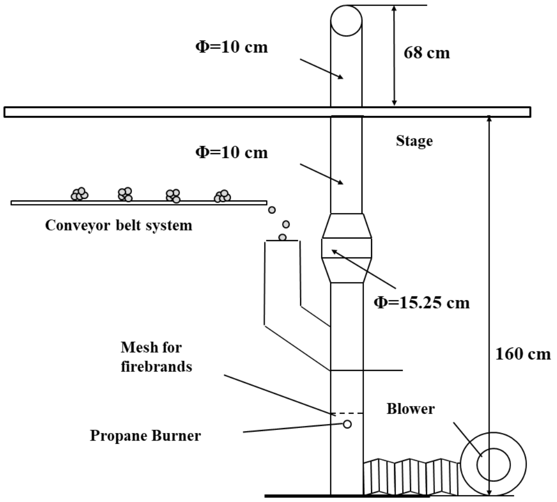


Figure 1 The reduced-scale firebrand generator is shown installed in NRIFD’s wind facility. Wood pieces are continuously feed into the device to generate firebrand showers.

The reduced-scale continuous-feed firebrand generator consisted of two parts; the main body and continuous feeding component. The capability of a smaller-sized firebrand generator to develop continuous firebrand showers has been described [**6**]. Japanese Cypress wood chips and Douglas-fir wood pieces were used to produce firebrands. The Japanese Cypress wood chips had dimensions of 28 mm ±7.5 mm (L) by 18 mm ± 6.3 mm (W) by 3 mm ± 0.8 mm (H) (average ± standard deviation), respectively, before combustion. These were provided from a supplier and filtered to remove really small wood chips by using a 1 cm mesh. Douglas-fir wood pieces were machined to dimensions of 7.9 mm (H) by 7.9 mm (W) by 12.7 mm (L).

The obstacle was placed downstream of the firebrand generator and the wind speed was varied at 4 m/s, 6 m/s, 8 m/s and 9 m/s. Specifically, the obstacle used in this study had the dimensions of 660 cm (H) by 1275 cm (W) and was located at a distance of 3 m from the device to visualize the transport process.

3. RESULTS AND DISCUSSION

Figure 2 displays images of experiments with a 660 cm (H) x 1275 cm (W) obstacle with different wind speeds (4 m/s and 8 m/s) and feeding materials (Japanese Cypress chips and Douglas-fir wood pieces). Images shows that firebrands made from Japanese Cypress wood chips and from Douglas-fir wood pieces accumulated differently in front of the same obstacle.

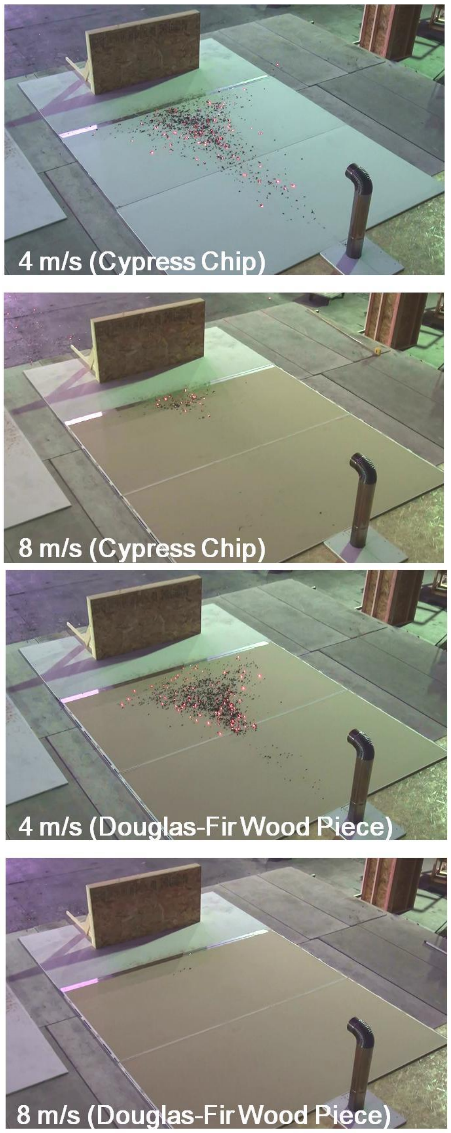

4 m/s (Cypress Chip)
8 m/s (Cypress Chip)
4 m/s (Douglas-Fir Wood Piece)
8 m/s (Douglas-Fir Wood Piece)

Figure 2 Images of experiments with a 660 cm (H) by 1275 cm (W) obstacle under different wind speeds -4 m/s and 8 m/s, with different feeding materials –Japanese Cypress wood chips and Douglas-fir wood pieces. 10 min of firebrand exposure has elapsed in these images.

Sub topic: Fire Research

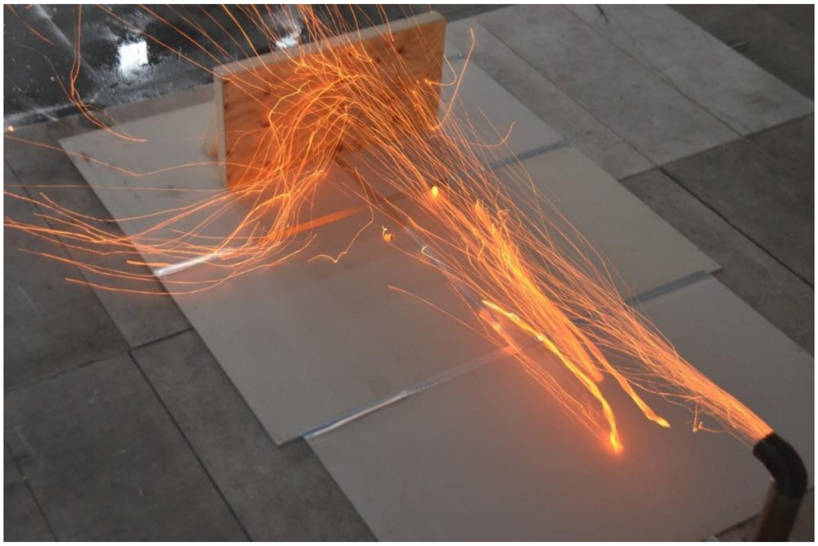


Figure 3 As an example of the physics involved, showers of laboratory generated firebrands combusting in a wind field of 8 m/s are directed at a simplified flow obstruction of 660 cm (H) by 1275 cm (W) obstacle. Here, Japanese Cypress wood chips are used as the source of firebrands.

Under a 4 m/s wind, it was observed that firebrands were unable to accumulate into one compact zone, rather scattered. As a wind speed increases, firebrands accumulated more in one zone. It was observed that firebrands made from Japanese Cypress wood chips managed to accumulate in front of the obstacle up to 10 m/s while those made from Douglas-fir wood pieces did not accumulate under 8 m/s and 10 m/s wind.

The accumulated area was measured using image processing software. Based on repeat measurements of different areas, the standard uncertainty in determining the projected area was ±10%. It is also important to mention that the obstacle was exposed to firebrand showers for the same duration. **Figure 3** displays an instantaneous image, displaying the complex physics involved.

Figure 4 displays the measured area of the accumulated firebrands as a function of wind speed. As the wind speed was increased, the area of accumulated firebrands was reduced significantly. It is obvious in **Figure 4** that firebrands from Douglas-fir wood pieces accumulated to smaller areas than those from Japanese Cypress wood chips.

Comparison was made in order to investigate the difference of accumulation behaviors of firebrands. **Figure 5** displays the mass and size distributions of firebrands made from Japanese Cypress wood chips and Douglas-fir wood pieces under the same wind speed (6 m/s). Similar to the authors prior full-scale studies, firebrands made from Japanese Cypress wood chips have approximately double the projected areas at a certain mass, compared to firebrands generated from Douglas-fir wood pieces. It suggests that the difference of firebrand characteristics has effect on the accumulation behavior of firebrands in front of wall assemblies.

Sub topic: Fire Research

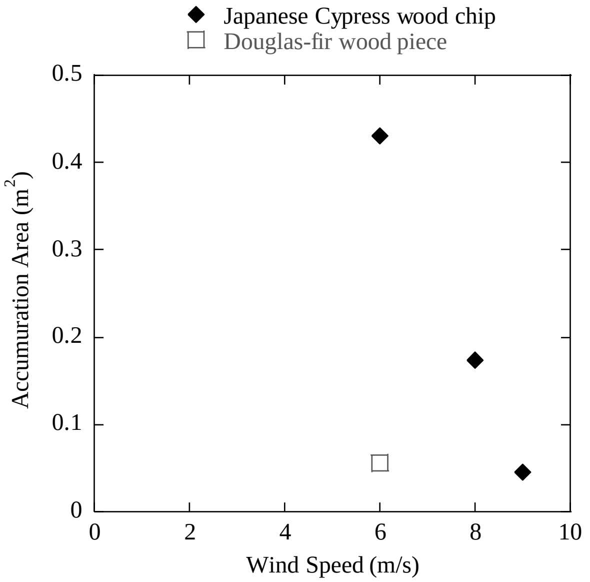


Figure 4 Measured accumulated area for various wind speeds.

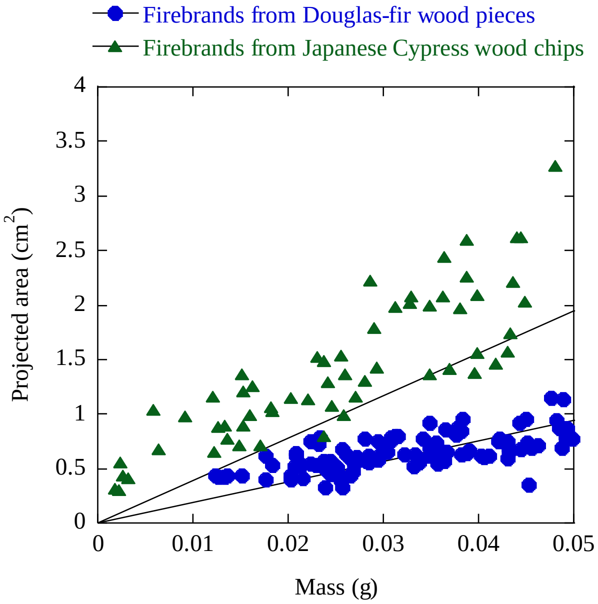


Figure 5 Mass and size distributions of firebrands made from Douglas-fir wood pieces and Japanese Cypress wood chips. Standard uncertainty in projected area (±10%) and mass (±1%).

4. COMPARISON OF ACCUMULATED INGITION PATTERNS

When firebrands accumulate into compact zones, it is known that they may provide sufficient heat feedback to ignite materials [**5**]. In full-scale experiments of Suzuki and Manzello

[**5**], when firebrands were observed to accumulate into compact zones in front of obstacles, the paper of the gypsum board was observed to ignite due to accumulated firebrands (see **Figure 6**).

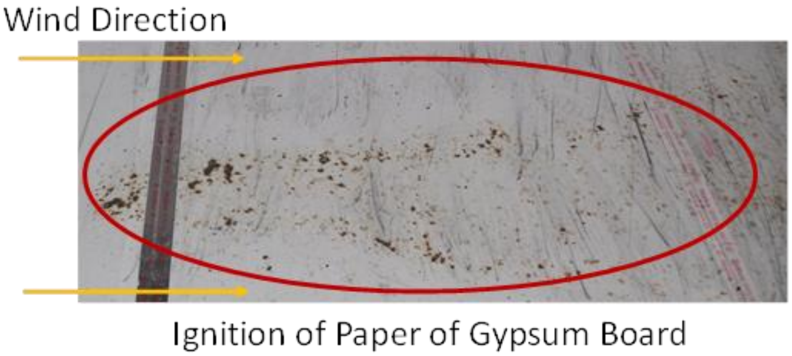


Figure 6 Accumulated firebrands swept off after the completion of the experiments. Ignition points of the gypsum board paper are observed. Douglas-fir wood pieces were used to simulate vegetative firebrands, wind speed of 6 m/s, and the obstacle was 1.32 m (H) by 2.44 (W) [**5**].

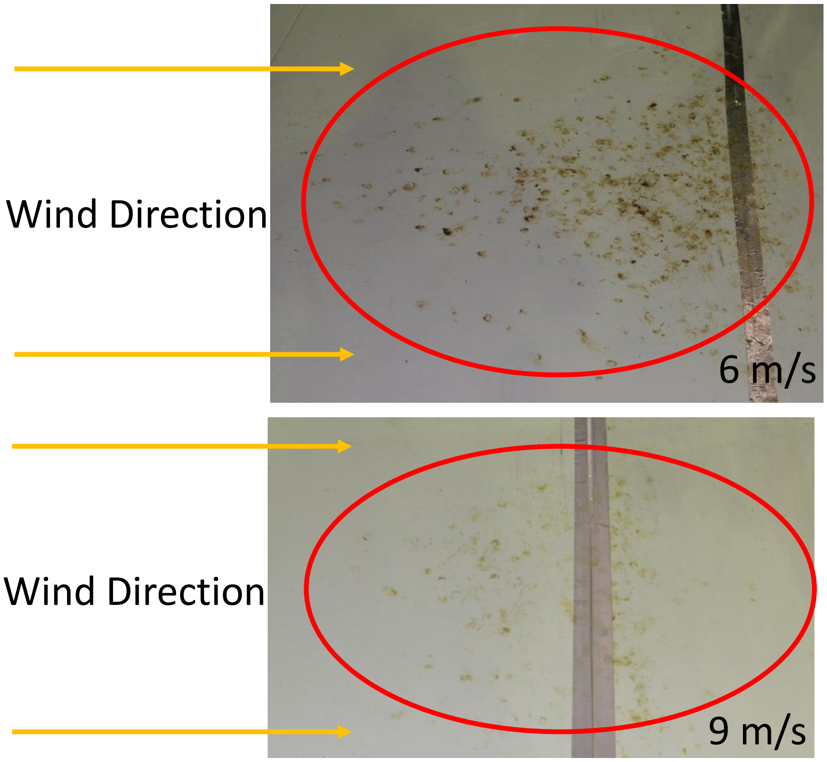


Figure 7 Accumulated firebrands swept off after the completion of the experiments. Ignition points of the gypsum board paper are observed at 6 m/s but not 9 m/s. Japanese Cypress wood chips were used to simulate structure firebrands, the wind speed was 6 m/s, and the obstacle was 660 cm (H) by 1275 cm (W).

Sub topic: Fire Research

Similar behavior was observed in these reduced-scale experiments. **Figure 7** displays images of the gypsum board surfaces after firebrands have been swept away. In the top image, the wind speed was 6 m/s and several ignition points were observed. At 9 m/s, it may be seen that, since the firebrands were unable to accumulate into compact zones, sufficient heat feed-back was not provided to ignite the gypsum board.

5. SUMMARY

A series of reduced-scale experiments were performed in order to investigate firebrand accumulation in front an obstacle using Japanese Cypress wood chips and Douglas-fir wood pieces. For a specified wind speed, it was found that the characteristics of firebrands has an effect of accumulation behavior of firebrands in front of an obstacle. Important qualitative similarities between the full-scale and reduced-scale firebrand studies is that increased wind speed resulted: (1) decreased firebrand accumulation areas (2) less ignition points on gypsum board surfaces. Detailed investigation will be needed in the future for different obstacles placed downwind from the firebrand generator, as well as further full-scale experiments generating firebrands similar to structure production for comparison. In addition, the authors plan to quantitatively measure the heat flux profiles imparted onto the surface under various experimental conditions. Methods to quantify heat flux profiles have been proposed and reviewed in the literature but it is not clear if these indeed work in more realistic experimental settings presented here [**2**].

6. REFERENCES

[1] S.L. Manzello, K. Almand, E. Guillaume, S. Vallerent, S. Hameury, and T. Hakkarainen, FORUM Position Paper, The Growing Global Wildland Urban Interface (WUI) Fire Dilemma: Priority Needs for Research, *Fire Safety Journal,* 100:64-66, 2018.

[2] S.L. Manzello, S. Suzuki, M. Gollner, and A.C. Fernandez-Pello, Role of Firebrand Combustion in Large Outdoor Fire Spread, *Progress in Energy and Combustion Science*, 76: 100801, 2020

[3] T, Iwami., et al., *Annual Meeting of Architectural Institute of Japan,* Tokyo, Japan, 2011. (in Japanese)

[4] S.L Manzello, S. Park, S. Suzuki, J. Shields, and Y. Hayashi, Experimental Investigation of Structure Vulnerabilities to Firebrand Showers, *Fire Safety Journal*, 46: 568-578, 2011.

[5] S. Suzuki and S.L. Manzello, Experimental Investigation of Firebrand Accumulation Zones in Front of Obstacles, *Fire Safety Journal*, 94: 1-7, 2017

[6] S. Suzuki and S.L. Manzello, Experiments to Provide the Scientific-Basis for Laboratory Standard Test Methods for Firebrand Exposure, *Fire Safety Journal*, 91:784-790, 2017.

NIST SP 1283
August 2022

Sub topic: Fire Research

12th U. S. National Combustion Meeting
Organized by the Central States Section of the Combustion Institute
May 24–26, 2021 (Virtual)
College Station, Texas

The Combustion of Noble-Fir Trees in the Presence of an Applied Wind Field

Samuel L. Manzello[1] and Sayaka Suzuki[2]

[1]National Institute of Standards and Technology (NIST), USA
[2]National Research Institute of Fire and Disaster (NRIFD), JAPAN
Corresponding author: samuelm@nist.gov

Abstract: Wildland fires that spread into urban areas, termed wildland-urban interface (WUI) fires, are becoming more and more common across multiple locations of the world. An important component in rapid spread of large outdoor fires is the production or generation of new, far smaller combustible fragments from the original fire source referred to as firebrands. Firebrands signifies any hot object in flight that are capable to ignite other fuel types. Firebrands are produced or generated from the combustion of vegetative and structural fuels. Firebrand processes include generation, transport, deposition, and ignition of various fuel types, leading to fire spread processes at distances far removed from the original fire source. The production of firebrands occurs from the combustion dynamics of vegetative and man-made fuel elements, such as homes. In this work, conifer trees (Noble-fir) were used to study the vegetative combustion process under an applied wind field. Two ignition methods were studied: the first employed a special propane burner and the second considered the use of firebrand showers impinging onto the trees. Specifically, temporally resolved mass loss profiles and heat flux profiles, as well as firebrand distributions were determined. Here, some initial findings of associated firebrand production are presented under 3 m/s and compared to experiments performed under no wind conditions. These experiments provide much needed experimental understanding needed to be able model vegetative combustion processes.

Keywords: large outdoor fires; firebrands; wind

1. INTRODUCTION

The large outdoor fire problem is a wide-spread global issue that shows no signs of abating. Perhaps the most well-known type of large outdoor fires are known as wildland fires that spread into developed, urban areas, known as wildland-urban interface (WUI) fires [**1-2**] As an example, the 2018 WUI fires in the US state of California demonstrated the shear destruction that WUI fires are capable of by destroying more than 18,800 structures and resulting in multiple fatalities [**3**]. In 2020, the situation grew even worse, with multiple fires reported all over the Western US. The size of areas burned in 2020 is simply staggering. In California, the August Complex fire itself consumed more than 1 million acres [**4**].

An important component in rapid spread of large outdoor fires is the production or generation of new, far smaller combustible fragments from the original fire source referred to as firebrands. Firebrands signifies any hot object in flight that are capable to ignite other fuel types. Firebrands are produced or generated from the combustion of vegetative and structural fuels. Firebrand processes include generation, transport, deposition, and ignition of various fuel types, leading to fire spread processes at distances far removed from the original fire source.

Manzello, Samuel; Suzuki, Sayaka. "The Combustion of Noble-Fir Trees in the Presence of an Applied Wind Field." Presented at 13th US National Combustion Meeting (13th US NCM). May 24, 2021 - May 26, 2021.

Sub topic: Fire Research

While it may be surprising to the reader, there has been little in the way of quantification of firebrands from vegetative fuel sources. Manzello and co-workers have recently completed a comprehensive review of firebrand processes [**5**]. To aid in understanding and to properly place this new study in the context of the past literature, some of this past work that has been reviewed has been repeated here for completeness.

In early pioneering work, laboratory experiments to investigate firebrands from actual tree combustion were performed by using Douglas-fir trees 5.2 m in height with a 3 m wide maximum girth by Manzello and co-workers [**6**]. As a first step, no wind was applied, and firebrands were collected by pans filled with water. Trees with 50% moisture content (MC) partially burned with no firebrands produced while the trees with 18% MC were engulfed in flames after only 20 s after ignition, producing numerous firebrands. Firebrands collected from trees with 18% MC had cylindrical shapes with an average size of 4 mm in diameter and a length of 53 mm. In addition, another experimental series was performed with Douglas-fir trees 2.4 m in height with a 1.5 m wide maximum girth. The average size of firebrands was 3 mm in diameter and a length of 40 mm.

In a follow-on study by Manzello *et al.*, [**7**], Korean pine was combusted at the Building Research Institute (BRI) Fire Research Wind Tunnel Facility (FRWTF) to investigate the difference in firebrand production from different tree species; no wind was applied. The height was kept constant at 4.0 m, and pans with water were placed around the tree for the firebrand collection. In order to have Korean pine combusted completely with a significant number of firebrands produced, it was found that MC had to be kept below 35% without any wind applied. The burn progressed somewhat sporadically, taking more than 2 min to complete. This was almost double duration for Douglas-fir trees (50–60 s). Firebrands were found to be cylindrical in shape with an average diameter of 5 mm and average length of 34 mm. The total mass of firebrands produced from each tree size was normalized with the mass lost from the tree during the burn as well as initial mass of the tree. While Douglas-fir trees showed a decrease in firebrand production (almost half) with an increase in tree height (almost double), Korean pine trees produced a larger ratio of mass of firebrands compared with Douglas-fir trees. With the ratio of burnable parts (needles and twigs) of Douglas-fir and Korean pine being similar, this reflects the difference of burning behavior between the two species. As mentioned previously, it took more than 2 min for Korean pine to burn completely while Douglas-fir trees burned out completely in 50–60 s for both heights tested. Douglas-fir trees also have a fuller, less open structure than Korean pine. The (HRR) estimated during the experiments in each case showed that, as for similar MC, Douglas- fir burns produced higher HRR than Korean pine. It was assumed that as most of firebrands produced or collected in this series were relatively small, the more intense fire plume from higher HRR Douglas fir burns might have consumed smaller firebrands completely before collection **[**7].

In more recent work, Blunck and co-workers [**8**] have been investigating firebrand combustion from various tree species. The most impressive component of their work is the sheer number of individual trees burned. Yet, none of the experiments were conducted under a controlled wind field, so the results are not a significant advancement to better understand firebrand generation processes in the present of wind [**8**].

In present study, Noble-fir trees were used to study the vegetative combustion process under an applied wind field. Noble-fir was used since it is easy to obtain from tree farms in the Western US and was easily imported to Japan. Two ignition methods were studied: the first employed a special propane burner and the second considered the use of firebrand showers impinging onto the trees. Specifically, temporally resolved mass loss profiles and heat flux profiles, as well as firebrand distributions were determined. Presently, some initial findings of associated firebrand production are presented under 3 m/s and compared to experiments performed under no wind conditions.

2. EXPERIMENTAL DESCRIPTION

A series of experiments were conducted using the wind facility at the National Research Institute of Fire and Disaster (NRIFD). The facility is able to generate a uniform wind profile with a cross section of 2.0 m by 2.0 m. The maximum wind speed possible using this facility is 10 m/s. The experiments made use of Noble-fir tree species, common in the Western US. The trees were imported from the state of Oregon in the USA. The maximum height of all trees was 1.5 m and was intentionally selected. The maximum girth or width varied as recorded for each experiment but was 1.0 m on average.

Upon arriving at NRIFD, the trees were stored in the laboratory to be able to control the MC. Samples of needles were taken on a daily basis, from three different areas of each tree, and oven dried to determine the MC. Similar procedures were used in prior, no wind studies [**6-7**]. A total of 20 trees were used for the experiments.

Each tree was placed on top of a load cell. Care was taken to protect the load cell from the heat generated from the combustion process. The trees were placed on top of gypsum board custom cut and sized to shield the load cells from the combustion process. The load cell was calibrated using a series of known mass samples prior to the experiments. The load cells were used to determine the temporal evolution of mass loss during combustion. In addition, a custom heat flux transducer array was fabricated for the experiments. Due to space considerations, these data are not presented here.,

To collect firebrands, both water pans as well as series of cameras were used to image firebrand production. Water pans were used to be able to directly compare results to legacy studies reviewed above.

A major goal of the present study was to determine an accepted ignition strategy for the trees under wind. As indicated, past experiments made use of no wind conditions to be able to gather basic understanding of the vegetative combustion process. Yet, it is obvious that in actual large outdoor fires, these combustion processes never occur without the present of wind. As a result, a new ignition strategy was needed. The most successful method made use of a 60 cm burner, oriented in a T shape, placed vertically along-side the trees.

Several burner configurations were attempted but this strategy was the most successful to yield sufficient combustion of the trees under wind. It was also observed that is was not possible to ignite the trees under no wind and then apply the wind. Such a strategy has been used by the authors for structural fuels, please see Suzuki and Manzello [**9**]. In the case of vegetative fuels, the subsequent ignition processes are very quick, on the order of tens of seconds, so by the time the wind was applied, nearly the entire tree would be consumed. Therefore, the strategy developed here allowed for ignition in the present of the applied wind field.

In addition to the T-shaped burner ignition methodology, there was interest to see if firebrand showers alone could result in tree ignition and combustion. For those experiments, the reduced-scale firebrand generator was used (see **Figure 1**). A reduced-scale continuous feed firebrand generator was used to generate firebrand showers. The description here closes follows Suzuki and Manzello [**10**] but is repeated here for completeness. This reduced-scale continuous feed firebrand generator consisted of two parts; the main body and continuous feeding component. The feeding part was connected to the main body and had two gates to prevent fire spread. Each gate was opened and closed alternatively. A blower was also connected to the main body and this was needed to loft and control the combustion state of the generated firebrands. The blower speed at the exit was set to 4.0 m/s to generate glowing firebrands. When the blower was set to provide an average velocity below 4.0 m/s, insufficient air was supplied for combustion and this resulted in a great deal of smoke being generated in addition to firebrands. 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, so these were generated. A longer main body was adapted so that only the firebrand generator part was above the stage, so the feeding part was not affected by wind. A conveyor was used to feed wood pieces continuously into the device. For all tests, Japanese Cypress wood chips were used to produce firebrands. These same size wood pieces have been shown to produce firebrands within projected area/mass of burning structures. As indicated, a conveyor was used to feed wood pieces continuously into the device. The conveyor belt was operated at 1.0 cm/s, and wood pieces were put on the conveyor belt at 12.5 cm intervals. The wood feed rate was fixed at 80 g/min, near the upper limit of reduced-scale firebrand generator.

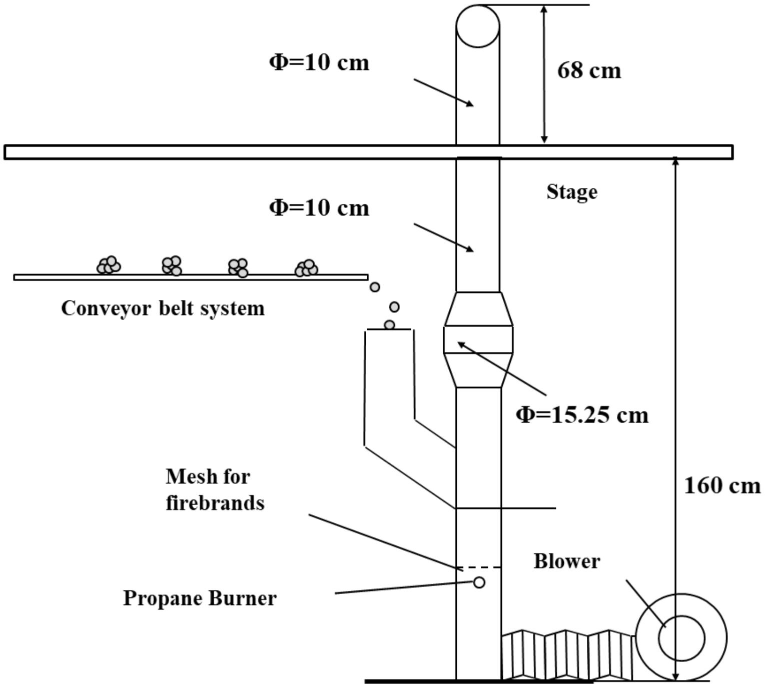


Figure 1 The reduced-scale firebrand generator is shown installed in NRIFD's wind facility. Wood pieces are continuously feed into the device to generate firebrand showers.

3. RESULTS AND DISCUSSION

Figure 2 displays the temporal evolution of tree ignition and subsequent combustion processes under a 3 m/s wind speed. The total tree height in this experiment was 1.5 m.

Sub topic: Fire Research

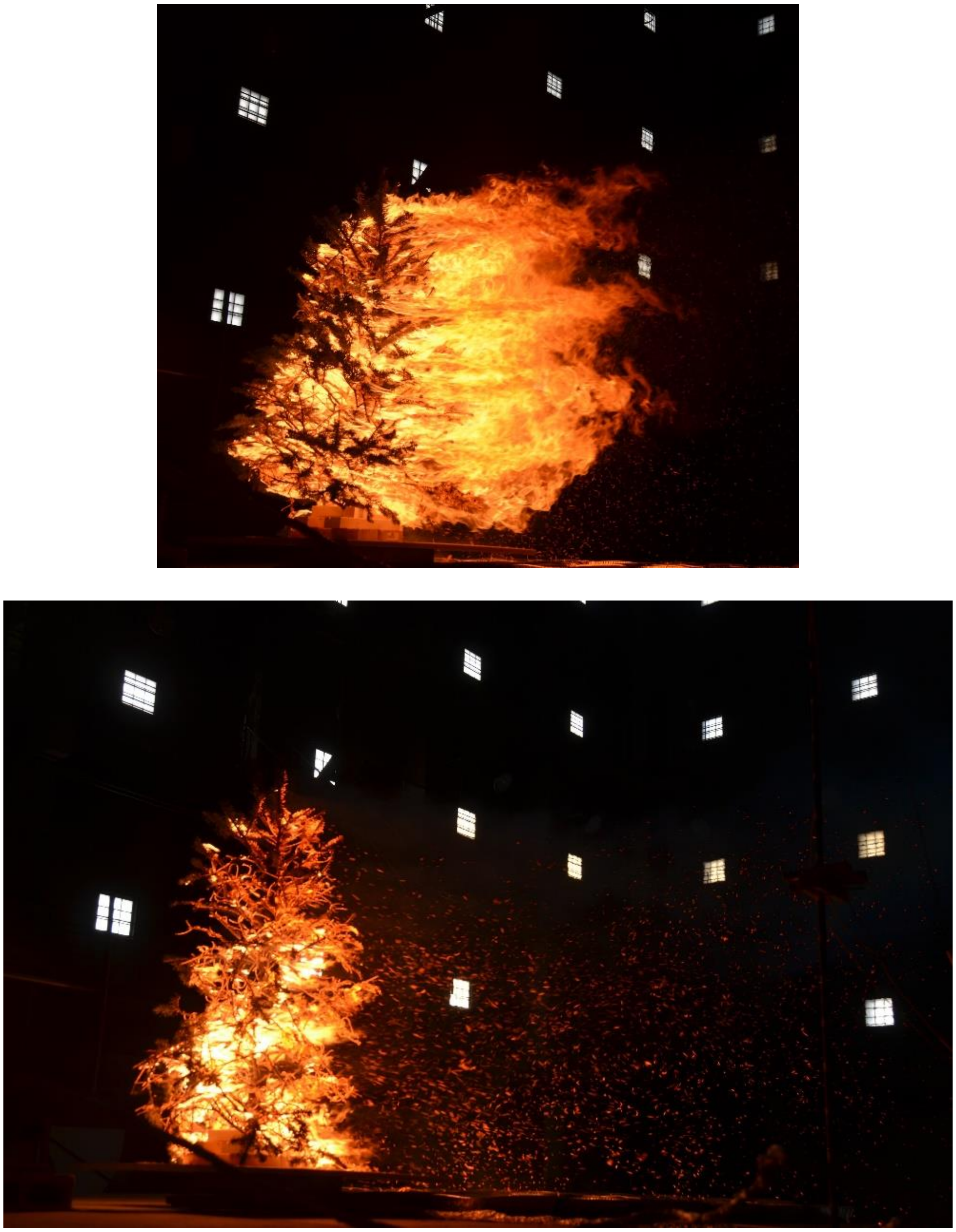


Figure 2 The top image shows the combustion process just after burner application and the bottom image shows the extensive firebrand generation processes. In this experiment, was applied wind speed was 3 m/s. The burner was applied for 10 s.

To clearly see the differences in studying the combustion processes under wind, another series of experiments were conducted under no wind. **Figure 3** displays a 1.5 m Noble-fir tree burning under no wind, with a similar MC as the tree in **Figure 2**. As can be seen, in the absence of wind, the buoyant fire plume remains vertical, as opposed to being titled on its axis in the present of wind.

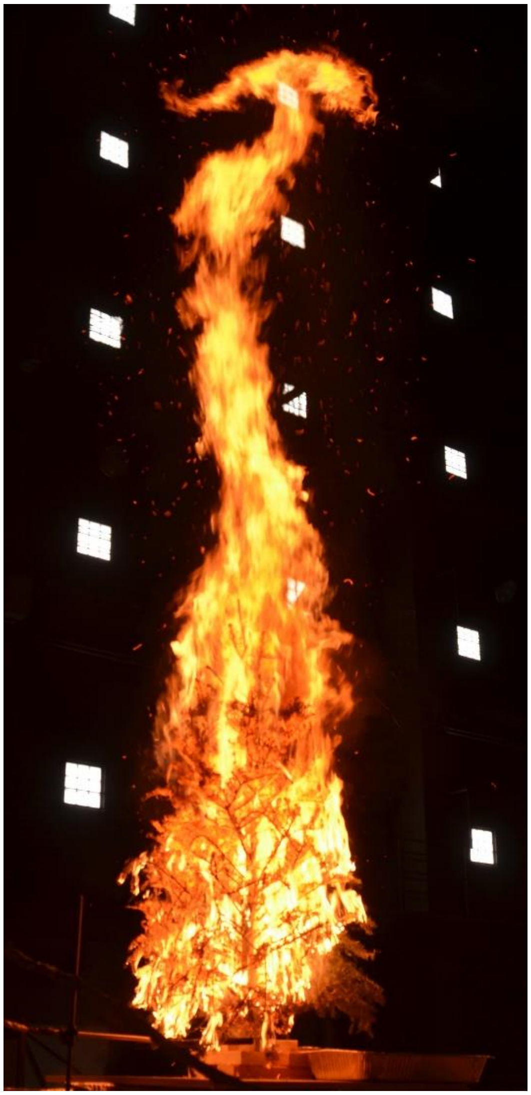


Figure 3 The combustion process just after burner application for a 1.5 m Noble-fir tree. In this experiment, no wind was applied. The differences are apparent in comparison to experiments under an applied wind field (see **Figure 2**). The burner was applied for 10 s.

3.1 Comparison to Ignition from Firebrand Showers

It was also desired to determine if it was possible to ignite trees using firebrand showers. In this case, the reduced scale firebrand generator was used. **Figure 4** displays the temporal evolution of the tree ignition process from firebrand showers at an applied wind speed of 3 m/s.

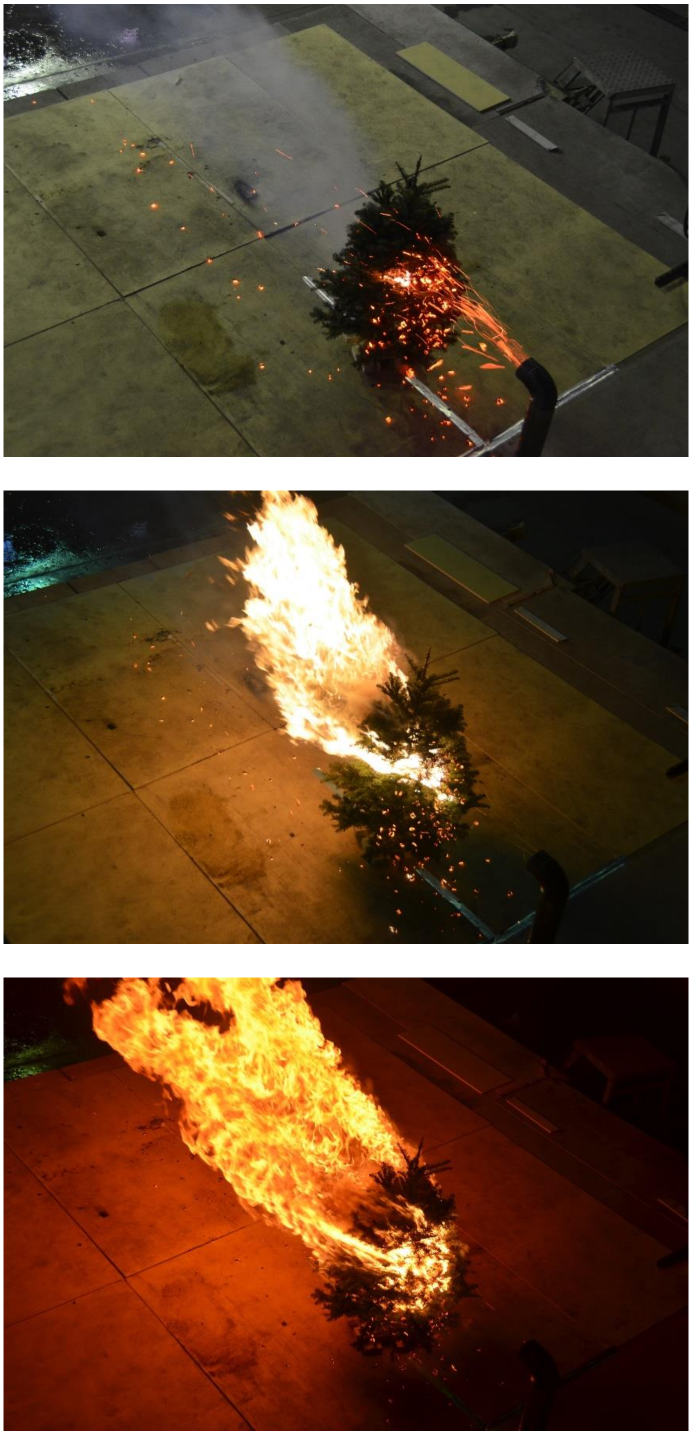


Figure 4 The combustion process of 1.5 m Noble-fir tree after ignition from firebrand showers. In this experiment, a 3 m/s wind was applied, and firebrand showers were selected to be commensurate to burning structures.

Sub topic: Fire Research

4. ASSOCIATED FIREBRAND PROCESSES

In the combustion process of vegetative fuels, pyrolysis of the fuel elements is an important mechanism. In the case of conifer tree combustion, fuel elements consist of needles, bark, and branches. During the combustion of vegetative fuel elements, these pyrolysis reactions result in the generation of gases and vapors and also act to weaken the structural integrity of the original fuel source itself as a result of the mass loss processes.

An interesting approach to model the generation of firebrands from vegetative fuels considered the use of fractal geometry to attempt to describe the various vegetative types [**11**]. In that work, comparisons were made to firebrand collected from Manzello *et al*., [**6**]; experiments of conifer tree combustion in the absence of wind.

As these combustion processes occur during actual large out fires, the application of aerodynamics forces from the interaction of wind forces imposed by the atmospheric boundary layer to vegetative fuel elements results in the breakage of small elements, that once lofted, become firebrands. **Figure 5** displays some of the collected firebrands from Noble-fir tree combustion, both under wind (3 m/s) as well as without wind application. As may be seen, the application of wind resulted in heavier firebrands being generated. Preliminary analysis of firebrand projected areas in shown in **Figure 6.** Standard uncertainty in projected area (±10%) and mass (±1%).

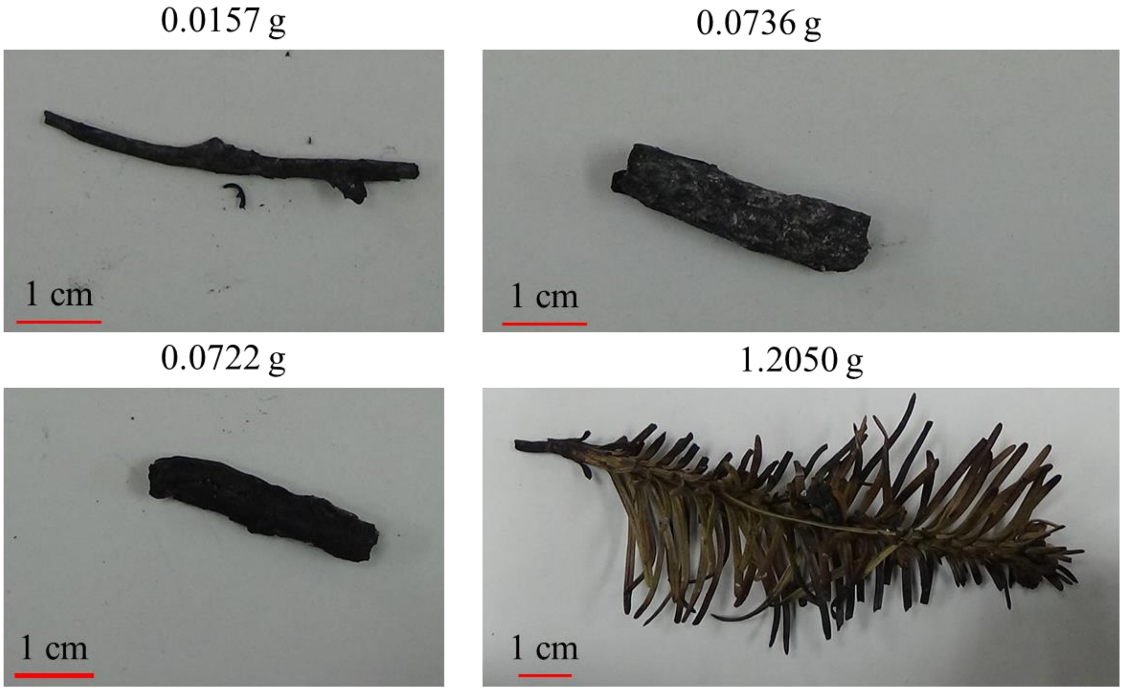


Figure 5 Samples of collected firebrands. The top panel shows firebrands collected from Noble-fir combustion under no wind. The bottom panel displays samples of firebrand collected under 3 m/s wind.

Sub topic: Fire Research

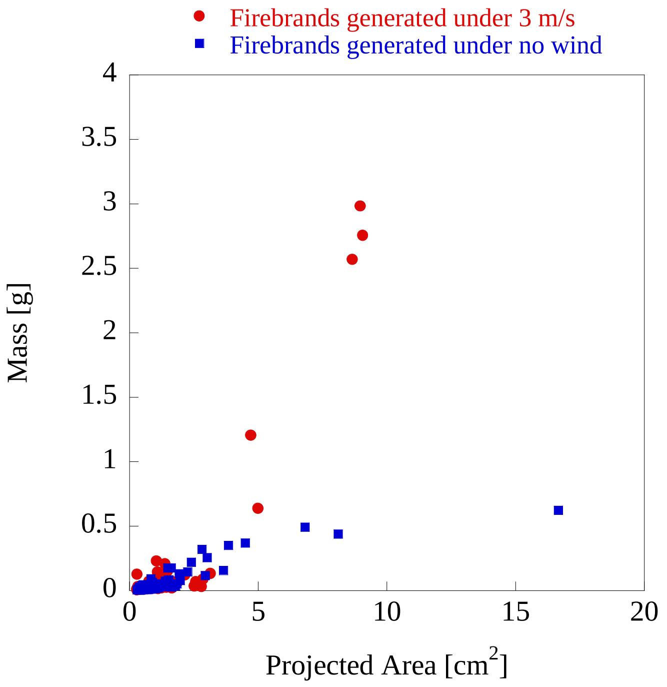


Figure 6 Mass and projected area of firebrands generated from the combustion of Noble-fir trees. Initial findings are presented for 3 m/s applied wind and compared to cases of no wind.

5. SUMMARY

In present study, Noble-fir trees were used to study the vegetative combustion process under an applied wind field. Noble-fir was used since it is easy to obtain from tree farms in the Western US and was easily imported to Japan. Two ignition methods were studied: the first employed a special propane burner and the second considered the use of firebrand showers impinging onto the trees. Some initial findings of associated firebrand production are presented under 3 m/s and compared to experiments performed under no wind conditions. This work presents an important step forward to understand vegetative combustion processes under wind.

6. REFERENCES

[1] S.L. Manzello, K. Almand, E. Guillaume, S. Vallerent, S. Hameury, and T. Hakkarainen, FORUM Position Paper, The Growing Global Wildland Urban Interface (WUI) Fire Dilemma: Priority Needs for Research, *Fire Safety Journal,* 100:64-66, 2018.

[2] A. Badia, M. Pallares-Barbera., N. Valldeperas, and M.Gisbert, Wildfires in the Wildland-Urban Interface in Catalonia: Vulnerability Analysis Based on Land Use and Land Cover Change. *Sci. Total Environ.*, 673, 184–196, 2019.

Sub topic: Fire Research

[3] S. Shulze *et al*., Wildfire Impacts on Schools and Hospitals following the 2018 Camp Fire *Natural Hazards*, 104: 901-925, 2020.

[4] California Department of Forestry and Fire Protection (CALFIERE), List Largest Wildfires Fires. https://www.fire.ca.gov/media/4jandlhh/top20_acres.pdf. Accessed March 30, 2021.

[5] S.L. Manzello, S. Suzuki, M. Gollner, and A.C. Fernandez-Pello, Role of Firebrand Combustion in Large Outdoor Fire Spread, *Progress in Energy and Combustion Science*, 76: 100801, 2020

[6] S.L. Manzello, A. Maranghides, and W. Mell, Firebrand Generation from Burning Vegetation*, Int'l J. Wildland Fire*, 16: 458-462, 2007.

[7] S.L. Manzello, A. Maranghides, J. R. Shields, W. E. Mell, Y. Hayashi, and D. Nii, Mass and Size Distribution of Firebrands Generated from Burning Korean Pine (*Pinus Koraiensis*) Trees, *Fire and Materials*, 33:21-31, 2009.

[8] T. Hudson *et al*., Effects of Fuel Morphology on Ember Generation Characteristics at the Tree scale, International J. Wildland Fire, 29: 1042-1051, 2020.

[9] S. Suzuki and S.L. Manzello, Garnering Understanding into Complex Firebrand Generation Processes from Large Outdoor Fires Using Simplistic Laboratory-Scale Experimental Methodologies, *Fuel*, 267, 117154, 2020.

[10] S. Suzuki and S.L. Manzello, Experiments to Provide the Scientific-Basis for Laboratory Standard Test Methods for Firebrand Exposure, *Fire Safety Journal*, 91:784-790, 2017.

[11] B.W. Barr, O.A. Ezekoye, O.A., Thermo-mechanical Modeling of Firebrand Breakage on a Fractal Tree *Proc. Combust. Inst. 34*: 2649-2656, 2013.

Transactive Energy and Solarization: Assessing the Potential for Demand Curve Management and Cost Savings

Himanshu Neema
himanshu.neema@vanderbilt.edu
Vanderbilt University
Nashville, TN, USA

Scott Phillips
Vanderbilt University
Nashville, TN, USA

Dasom Lee
University of Twente
Enschede, Netherlands

David J. Hess
Vanderbilt University
Nashville, TN, USA

Zachariah Threet
Tennessee Technological University
Cookeville, TN, USA

Thomas Roth
Cuong Nguyen
National Institute of
Standards & Technology
Gaithersburg, MD, USA

ABSTRACT

Utilities and local power providers throughout the world have recognized the advantages of the "smart grid" to encourage consumers to engage in greater energy efficiency. The digitalization of electricity and the consumer interface enables utilities to develop pricing arrangements that can smooth peak load. Time-varying price signals can enable devices associated with heating, air conditioning, and ventilation (HVAC) systems to communicate with market prices in order to more efficiently configure energy demand. Moreover, the shorter time intervals and greater collection of data can facilitate the integration of distributed renewable energy into the power grid. This study contributes to the understanding of time-varying pricing using a model that examines the extent to which transactive energy can reduce economic costs of an aggregated group of households with varying levels of distributed solar energy. It also considers the potential for transactive energy to smooth the demand curve.

CCS CONCEPTS

• **Computer systems organization → Embedded and cyber-physical systems**; • **General and reference** → *Design*; • **Computing methodologies → Model development and analysis**; • **Applied computing → Engineering**.

KEYWORDS

Transactive Energy, Community Choice Aggregation, Virtual Power Plant, Modeling and Simulation, Cyber-Physical Systems, Societal Implications

ACM Reference Format:
Himanshu Neema, Scott Phillips, Dasom Lee, David J. Hess, Zachariah Threet, Thomas Roth, and Cuong Nguyen. 2021. Transactive Energy and Solarization: Assessing the Potential for Demand Curve Management and Cost Savings. In *Design Automation for CPS and IoT (Destion '21), May 18, 2021, Nashville, TN, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3445034.3460510

1 INTRODUCTION

This study examines the interaction of transactive energy with distributed generation for a localized community of consumers who also produce some of their own electricity. In other words, it creates a virtual power plant (VPP) that is managed by pricing signals and adopts widely discussed parameters (variables), such as pricing techniques, battery presence, distributed energy resources (DERs), and load management through remote adjustment of appliances such as air conditioning units.

The key contributions of this study are twofold. First, it tests the effectiveness of pricing techniques to control load management. The findings of this paper can be directly applied in real life settings as it uses real energy consumption and weather data in Sacramento, California (CA). Second, by adopting several parameters, such as battery use, solar penetration rate, wattage of solar, and pre-cooling systems, this study attempts to produce the most effective transactive energy model that adopts solar energy. In this sense, the findings can lead to policy proposals that aid the development of transactive energy systems. It is especially applicable for regions where the solar penetration rate is high and increasing.

Transactive energy (TE) refers to *the combination of economic and control techniques to improve grid reliability and efficiency* [15]. One of the fundamental goals of TE is to coordinate the operation of new energy systems that contribute to the efficiency and reliability of the grid, which include many DERs. DERs refer to renewable energy generation technologies deployed in the distribution grid [13] at consumer's side and include local storage batteries, roof-top solar photovoltaic units, wind-generation units, and biomass generators. This study specifically focuses on the adoption of distributed solar energy and battery presence.

The study uses energy and pricing generation data from Sacramento, CA. In general, the state has government policies with relatively advanced use of time-of-use pricing and distributed energy generation. Customers in the state are served by public power companies for some cities (e.g., Sacramento, CA), rural electricity cooperatives, investor-owned utilities, and increasingly community-choice aggregation (CCA) organizations. The state does not have

Neema, Himanshu; Phillips, Scott; Lee, Dasom; Hess, David; Threet, Zachariah; Roth, Thomas; Nguyen, Cuong. "Transactive Energy and Solarization: Assessing the Potential for Demand Curve Management and Cost Savings." Presented at Design Automation for CPS and IoT (DESTION 2021), Nashville, TN, US. May 18, 2021 - May 18, 2021.

retail competition for customers, but CCAs in California have grown rapidly [12] [21]. Customer aggregation occurs when a local government or group of local governments enroll customers in their jurisdiction to purchase electricity for them as a collective unit. Doing so can provide customers with better prices. In California, some CCA organizations developed or launched in a form where the local government is involved in managing contracts and supporting energy efficiency and renewable energy generation, and the CCA organization begins to approximate a public power agency [20]. In these more advanced forms of CCAs, there is growing interest in VPPs. This research focuses on examining various DER configurations in VPPs.

Although the data and context are based on California, the analysis is relevant for other electricity utilities, cooperatives, and public power organizations in other states or countries that are considering the use of VPPs to enhance grid stability and efficiency. Any electricity provider that also wants to integrate higher levels of distributed renewable energy with pricing programs for demand management would find the analysis to be relevant. The study generates experiments to model the effects of TE on energy consumption decisions with varying levels of distributed renewable energy and examines the effects on household costs.

The central research question is: *what are the most efficient parameters in a decentralized energy system?* In answering this question, several relevant parameters are considered, such as pricing techniques, battery presence, the penetration rate of DERs, and load management and pricing optimization techniques such as precooling. The two pricing techniques tested were *time-of-use pricing* (TOU) and *real-time pricing* (RTP). In TOU, the price follows a set schedule, generally changing a few times throughout the day. In RTP, the price is varied over very short time intervals based on projected demand in those time frames. The experimental results show relationships between various combinations of these parameters and their effects on smoothing of the demand curve.

2 MOTIVATION

The use of DERs for energy generation has been increasing in California, which has the highest installed capacity of DERs in the United States with 3154 MW of DERs installed in 2014 [13]. However, despite their popularity, installation and maintenance cost is deterring many potential consumers from transitioning to DERs. A simulation model that indicates the economic savings of DER adoption and the most effective DER model could encourage more consumers to adopt DERs.

In addition, emerging TE technologies could smooth the demand curve and provide energy price systems that adhere to the supply and demand of energy at any given time [8] [17]. These benefits are expected to lead to more efficient grid management and cost savings. This study will help to clarify how different real-time pricing configurations interact with changes in the grid configurations, including DERs and energy storage.

By directly linking DER and TE and by using real-world data, this study builds on and complements experimental projects such as [11] [25] [14] by exploring a wider range of potential configurations of parameters than those are generally found in these projects.

Power grids equipped with TE are a highly complex CPS with additional human and economic factors that must be considered. A systematic approach to automate design of RTP models can lead to more adaptive load management as well as operational efficiencies.

3 RELATED WORK

3.1 Distributed Energy Resources

The role of distributed energy resources (DERs) has become increasingly important, particularly as the shift towards renewable energy becomes more prevalent. Concurrent with the growth of DER is the development of energy storage technologies, such as battery energy storage systems and flywheels [4].

DERs can be configured as microgrids and VPPs for enhancing their efficiency, cost effectiveness, and resilience [19]. In particular, VPPs are considered to be a cost-efficient integration of DERs [6]. This study contributes to the existing literature by examining various configurations of DER in a VPP setting that uses TE.

3.2 Transactive Energy

Transactive energy (TE) has been widely discussed in the literature, and it is considered as one of the key energy technologies that will result in more renewable and sustainable energy production and consumption. The advantages of TE are largely twofold: first, it allows for more predictive energy demand and consequently more efficient load management [9]. For example, [22] argued that buildings are 70 % of the load on today's grid, which makes the shape of electricity loads critically important in improving energy efficiency. TE can play an important role in smoothing the daily load. Second, for various reasons (more efficient load, integration of DER, etc.), transactive energy is also expected to dramatically reduce energy costs for both consumers and producers [9]. Transactive systems have been shown to reduce expected costs up to 75 % when local markets and flexibility are considered [18].

In this sense, real-time pricing (RTP) is one of the key characteristics of TE [8]. RTP requires constant data collection, often as frequent as every 5 minutes, which allows for more efficient peak demand and use of energy [27]. A 5 % reduction of peak demand can lead to $3 billion savings in the United States [10]. Furthermore, peak load reductions can lead to environmental benefits and a reduction of emissions. On the consumer side, there are economic benefits of RTP. Zethmayr and Kolata found that 97 % of regional customers would have saved money with RTP without changing their behavior because flat-rate supply service tends to incur higher bills than the hourly market price [27].

However, there are also challenges associated with RTP. One of the biggest challenges is privacy because there is a constant collection of energy consumption data. In order to solve the problem, there is ongoing innovation in privacy standards, guidelines, and regulation in various countries [16].

One of the most commonly used implementation methods for price modification is transactive incentive signals, in which the signal is sent by the utility to each consumer. Another commonly used implementation method is the transactive feedback signal, which is sent from the consumer to the utility and contains information regarding expected energy use. This has been modeled as time-of-use-based demand-response [24].

4 EXPERIMENT DESIGN

In order to simulate the community, their energy use, and the effect of different parameters, the system is modeled in a widely used open-source power-distribution system simulator known as GridLAB-D [7].

The simulation is based in Sacramento, CA, where DERs are widely prevalent and the region has sufficient weather variations to demonstrate dynamic feedback and control. The simulation models a community with 544 residential houses and 26 businesses using 2017 weather data, which was the latest and most reliable weather data for that area [3]. In order to optimize solar DER, a number of parameters are employed, such as solar panel power, solar panel penetration, the presence or absence of batteries, and pre-cooling. All simulations were run from 0:00:00 8/1/2017 to 0:00:00 8/15/2017. No heating was added to the model because it is unlikely that heating will be used in this area during the summer periods. Running the same simulation in the winter months likely would have resulted in solar panels and batteries used to store solar energy for later use and in having a reduced ability to flatten the load curve.

The above parameters were varied to observe their effects on the relevant outcomes: reducing utility demand, cutting costs for communities, reducing peak load on the system, and flattening the load curve. Each parameter combination was paired with both the TOU and RTP price schedules.

The functionality through which these outcomes could be improved is demand response. In other words, consumers will alter their energy demand based on changes in price. This is included in the model in the HVAC units. The behavior of these units is modeled by passive controllers in GridLAB-D, which, according to the procedure described in the following section, will decrease air conditioning use when prices are high.

4.1 Data Collection

Data were collected using government websites that published energy data, independent solar companies' websites, and other published data. Simulating a model requires assumptions on energy consumption, the choice of solar panels (i.e., efficiency of the solar panel), battery storage, and energy pricing scheme. Consequently, the following assumptions were made based on the data collected. Each household in California consumes approximately 546 kWh per month. This assumption is based on the 2018 data published by the U.S. Energy Information Administration [1]. This is the most recent dataset available for energy consumption per state, and there was little change in energy consumption in California from 2016 and 2018. Therefore, it is assumed that there is little change in energy consumption from 2018 to 2020.

Solar panels tend to vary in prices and efficiency. The five most popular solar panels in California (Solar Estimate 2020) were averaged, which makes the efficiency of the simulated solar panels 19.52 %. The average size of home solar installations in California in 2019 was 7 kW.

4.2 Experiment Parameters

Pricing technique: One of the key characteristics of TE is the adoption of different pricing techniques. This study adopts two different pricing techniques: time-of-use (TOU) and real-time pricing (RTP). TOU is one of the most widely used pricing techniques currently used in California. RTP is currently not deployed because of regulatory limitations regarding energy data collection, but it is the most commonly used pricing structure in TE [26]. Therefore, using an alternative platform such as simulation modeling is particularly useful in comparing these two pricing techniques. The two pricing systems are explained in detail in Section 5.

Wattage of the solar system: Simulations were executed with 5 kW solar panels, a common residential solar system generation capacity, and 7 kW, the most common generation capacity of residential installations in California.

Battery presence: The simulation also considers whether battery presence changes the efficiency of DERs. Batteries are considered particularly useful in many solar systems because they can provide energy after the sunset or even during days when direct sunlight is not available due to the weather restrictions. Each household is assumed to contain one battery storage system modeled on Tesla Powerwall, which includes a built-in inverter. However, the battery model can be configured for different capacities. One battery is assumed to be enough because it not only simplifies modeling, but, more importantly, it also removes the constraint of using two batteries at all houses, which could be cost prohibitive for some houses. Moreover, a single battery with configurable capacity makes the study more adaptable and flexible under different conditions and for different areas.

Solar penetration rate: The simulation introduces variation in the solar penetration rate of the community to analyze the extent to which solar adoption affects the demand curve, and consequently, grid management and reliability. Therefore, three different solar penetration rates are tested: 0 %, 25 %, and 50 %.

Pre-cooling: Without pre-cooling, the behavior of HVAC systems is determined only by the current price of power. With pre-cooling enabled, HVAC systems will cool houses in advance of future power price increases, with the goal of saving money for consumers. Both conditions are tested.

5 SYSTEM ARCHITECTURE

In this section, we describe our approach to modeling the power grid and load, TOU and RTP pricing, and experiment automation.

5.1 Modeling the Grid and Load

The GridLAB-D model is based on the feeder model R1-12.47-2 [23] developed by Pacific Northwest National Laboratory (PNNL). The feeder model is comprised of a moderately populated suburban and sparsely populated rural area in which approximately 70 % of the circuit-feet are overhead and 30 % are underground. This feeder model was extended by adding triplex meters connecting existing triplex nodes to the houses. Triplex meters were also added to connect each solar panel and inverter pair to the grid (see Figure 1). In addition to power consumption from the HVAC, each residential house was connected to two ziploads: one using GridLAB-D's built-in *unresponsive_load* schedule, and the other using the built-in *responsive_load* schedule. The power use for the 26 commercial entities in the simulation was setup with unresponsive ziploads. Each business had loads modeling interior lighting, exterior lighting, plugs, gas waterheater, and occupancy.

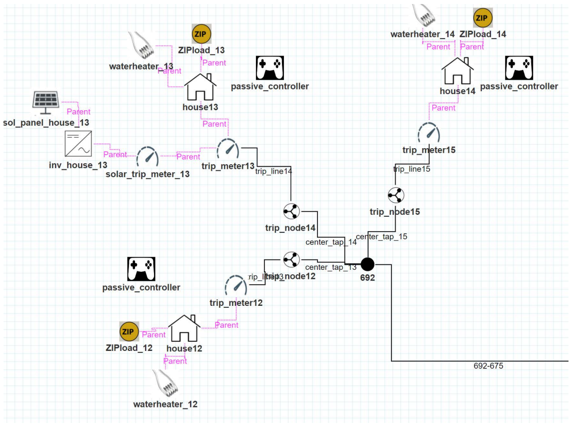


Figure 1: Part of the Grid Model used in Experiments

Consumer response to price changes is primarily modeled by a passive controller connected to each residential home's HVAC system. All HVAC's in our model were electric. These passive controllers respond to the price changes in the *stubauction* object (a basic GridLAB-D auction module), determined by the price schedule being used. The parameters *range_low* and *range_high* represent the most the consumer is willing for their house temperature to change due to transactive control. The range_low parameter was set as a random number in the range between -1 and -2, and range_high was set at a random number between 2 and 4. This means that the consumer with average preferences (76 °F base temperature) would be willing to have their HVAC vary the temperature between about 74 °F and 80 °F.

The stubauction object in GridLAB-D was used in conjunction with passive controllers, which responds to price changes. The stubauction period was 300 s, meaning that RTP price updates occur every 5 minutes.

Within the maximum range of temperature alterations, HVAC behavior is modeled by piece-wise functions [2] corresponding to customer willingness to change temperature within their maximum range. Variable *ramp_high* represents the temperature increase that would accompany a 1 standard deviation increase in price. Variable *ramp_low* represents the temperature decrease that a customer would be willing to pre-cool to in the case of a price 1 standard deviation below average. Both ramp_high and ramp_low are set to a random number between 1 and 4.

Group IDs were assigned to each triplex meter indicating whether it was residential, commercial or solar meter. Using these group IDs, collectors aggregate the data from each of residential, commercial, and solar entities to get a clear picture of energy production and consumption in each simulation. The default powerflow solver method in GridLAB-D, Forward Backward Sweep (FBS), was used for the simulation. The minimum timestep in GridLAB-D was set to 15 s, which was found to offer enough precision without resulting in prohibitively long simulation runtimes. Treatment of transient stability would require enforcing transient stability thresholds as invariants and relating transient stability with price variations, which is outside the scope of this paper.

The temperature inside the house can greatly affect the energy bill. The average cooling temperature in hot-dry states, which includes California, is 76.4 °F [5]. The cooling set points were randomized for each house using a uniform distribution within 2 °F of this average, between 74 °F to 78 °F. This specific range was chosen to give variation among different households, which would result in a more realistic data.

5.2 TOU & RTP Pricing

For TOU, data were used from Marin County, CA. During weekdays, the peak demand period is from 1 p.m. to 7 p.m. and park peak period is from 10 a.m. to 1 p.m. as well as 7 p.m. to 9 p.m. During weekends, the park peak period is from 5 p.m. to 8 p.m. All other hours are considered off-peak hours.

RTP price schedules are generated from the results of TOU simulations. Two strategies were employed to reduce peak power usage and flatten the demand curve: raising prices during times of high usage and lowering prices in advance of high usage periods, both based on the TOU simulation results. Two *lookahead periods* were used in which the demand over the next set of time-frames would be averaged together. In the first, shorter lookahead period, higher demand in the TOU simulation corresponds to proportionally higher price for the RTP price schedule. In the second, longer lookahead period, higher demand in the TOU simulation would correspond to a proportionally lower price for the RTP price schedule. These two candidate price schedules are averaged together according to a weighting.

Based on the TOU simulation results, let U_p be the mean power usage in the previous *'x'* time-slots of 5 minutes each, and U_n be the mean power usage in the next *'y'* time-slots of 5 minutes each, A be the average power usage for the entire simulation, and *w1* and *w2* be the weighting of the *'x'* and *'y'* lookahead periods respectively (where, *w1* + *w2* = 1), then

$$P_{+} = U_p/A * avg\_price \tag{1}$$

$$P_{-} = (2 - A/U_n) * avg\_price \tag{2}$$

$$P_f = P_{+} * w1 + P_{-} * w2 \tag{3}$$

After setting up TOU and RTP capabilities, parameters x, y, and *w1* (as *w2* = 1 - *w1*) were varied to determine the best combination for an RTP price schedule. The process of generating price schedules, running the GridLAB-D simulation, and processing the results of the simulation were automated with a bash script. After each simulation run, results such as residential load, commercial load, and solar load were reported every 5 minutes. From these results, a Python script automatically calculated metrics such as the peak power demand, the standard deviation of power demand, and a statistic we created: a 6-hours $MaxMin_d$. In $MaxMin_d$ metric, the average of 5 minimum load time-frames was subtracted from the average of 5 maximum load time-frames over each 6-hours time-period of the simulation. This metric aimed to show which simulation parameters resulted in large variation over short time-frames, even if the absolute peak load over the simulation wasn't extremely high.

Name	Max (VA)	SD (VA)	$MaxMin_d(VA)$	Res Bill ($)	Sol Bill ($)
7_batt_RTP_50	2070750	306491	2113857	65.09	-62.94
7_batt_RTP_50_nopre	2179040	324463	2142096	64.18	-61.71
5_batt_RTP_50	2464840	358899	2389285	74.73	-15.67
5_batt_RTP_50_nopre	2501670	377959	2403859	73.96	-14.88
7_batt_RTP_25	2469690	425150	2653535	82.51	22.98
7_batt_RTP_25_nopre	2460750	438485	2658196	82.15	22.94
5_batt_RTP_25	2766350	457178	2834549	87.24	45.01
5_batt_RTP_25_nopre	2700960	470534	2872458	86.98	44.77
5_nobatt_RTP_50	2538620	419921	3036072	74.75	-7.45
7_batt_TOU_50	3188990	348845	3210814	69.55	-81.64
7_batt_TOU_50_nopre	3185220	347426	3219136	69.35	-81.84
7_nobatt_RTP_50	2485180	436913	3340449	65.55	-44.11
0_batt_RTP_50	2860000	560933	3399355	99.30	99.30
0_batt_RTP_25	2816270	562452	3416696	98.86	98.86
0_nobatt_RTP_50	2901270	559742	3436602	99.06	99.06
0_batt_RTP_50_nopre	2883640	577873	3508441	98.43	98.43
0_batt_RTP_25_nopre	2986200	578469	3538669	98.76	98.76
5_batt_TOU_50	3550930	403705	3638344	80.60	-27.39
5_batt_TOU_50_nopre	3527530	404007	3671989	80.47	-27.52
7_batt_TOU_25	3461310	467367	4025024	89.99	18.35
7_batt_TOU_25_nopre	3453180	468556	4061837	89.80	18.16
5_batt_TOU_25	3591500	499584	4356122	95.25	44.08
5_batt_TOU_25_nopre	3621350	500911	4377977	95.13	43.96
7_nobatt_TOU_25	3626040	500120	4480413	90.68	19.04
5_nobatt_TOU_50	3619740	481036	4493945	81.75	-26.24
5_nobatt_TOU_25	3642000	522936	4561583	95.79	44.62
7_nobatt_TOU_50	3656030	496599	4816567	70.99	-80.20
0_batt_TOU_50	3637330	611532	5055585	107.96	107.96
0_nobatt_TOU_50	3659540	609829	5064896	108.57	108.57
0_nobatt_TOU_25	3640000	609522	5070649	108.50	108.50
0_batt_TOU_25	3649990	610658	5081957	108.31	108.31
0_batt_TOU_25_nopre	3652230	611289	5098970	107.96	107.96
0_batt_TOU_50_nopre	3645750	613297	5128058	107.84	107.84

Table 1. Simulation Results

The values of $x = 6$ time-slots (30 minutes), $y = 115$ time-slots (575 minutes), *w1* = 0.1, and *w2* = 0.9 were found to minimize peak load, standard deviation, and 6-hours $MaxMin_d$.

5.3 Experiment Automation

In order to run large numbers of experiments efficiently, the process of running GridLAB-D simulations was automated using a bash script (see Figure 2). The simulations are run by entering the following line (or a variation) in git bash:

sh runGLD.sh panel_power battery <.GLM file> x y w1

Here, *panel_power* is an integer representing the power in kW of the solar panels; *battery* specifies running a simulation with 13.5 kW batteries, while the alternative *nobattery* specifies running a simulation with no batteries present.

First, the bash script edits the price schedule generation Python script using the desired parameters: *x*, *y*, and *w1.* The price schedule generation script then executes, using these parameters and a previous TOU simulation to generate the new price schedule. Next the bash script edits the GridLAB-D model file according to the specified parameters. After the simulation, the bash script generates the bills for residential customers with solar panels, residential customers without solar panels, and commercial entities, as well as calculating descriptive statistics of the power load graph.

6 EXPERIMENT RESULTS

Table 1 shows a subset of the simulations run along with some evaluation metrics. The name of the simulation is comprised of the following parameters in order:

- Power generation capacity of the solar panels (0 kWh, 5 kWh, or 7 kWh);
- *batt* if the simulations have batteries present, otherwise *nobatt*;
- *RTP* if real-time pricing was used, and *TOU* if time of use pricing was used;
- The percentage of residential houses that have solar panels (25% or 50%); and
- *nopre* if precooling was disabled for that simulation.

The column names in the simulation results, from left to right, shows: name of the simulation (*Name*); grid power load (*Max*); the standard deviation of grid power (*SD*); the differences between maximum and minimum ($MaxMin_d$ – defined in Section 5); the power bill for residential customers without solar panels for the 2 weeks simulation (*ResBill*); and the power bill for residential customers with solar panels for the 2 weeks simulation (*SolBill*).

6.1 Effects of Parameters on Power Costs

There were three main effects of the parameters on power costs:

(1) **RTP vs TOU pricing**: The RTP model used for this study does not seem to drastically reduce prices for consumers. We observed that in each RTP vs TOU simulation pair in which the other variables were held constant, RTP slightly reduces the energy bill for residents without solar panels, and increases the energy bill for residents with solar panels. One potential reason for this could be shifting of slightly higher prices to price responsive customers.

(2) **Battery storage**: Batteries are found to have little effect on the energy bill of residents without solar panels, but reduce the bill for residents with solar panels. A potential reason for this could be that for customers without solar panels, battery could serve to store power during lower prices, which could later be used when prices rise. On the other hand, for customers with solar panels, the effect of local storage might be negated by the higher prices they could get by pushing excess power in the grid during that time.

(3) **Solar panel penetration and generation capacity**: Higher solar panel penetration and higher solar panel power generation capacity both correspond with significant reductions in the energy bill for both categories of customers due to overall reduction in power demand from utilities.

6.2 Effects of Parameters on the Duck Curve

There were also significant impact of the parameters on the *duck curve* (a graph of grid power load that dips in the middle of the day during solar power generation and then rises at the end of the day as people use more power in the evenings) as follows:

(1) **RTP vs TOU pricing**: RTP significantly smoothed the duck curve when compared to TOU by shifting load before peak usage (see Figure 3). RTP simulations have lower *max* power, standard deviation (*SD*), and $MaxMin_d$ than TOU simulations when other variables are kept constant. In fact, TOU simulations with transactive control, had more undesirable duck curves than without TE, potentially due to abrupt changes in power demand when price changes.

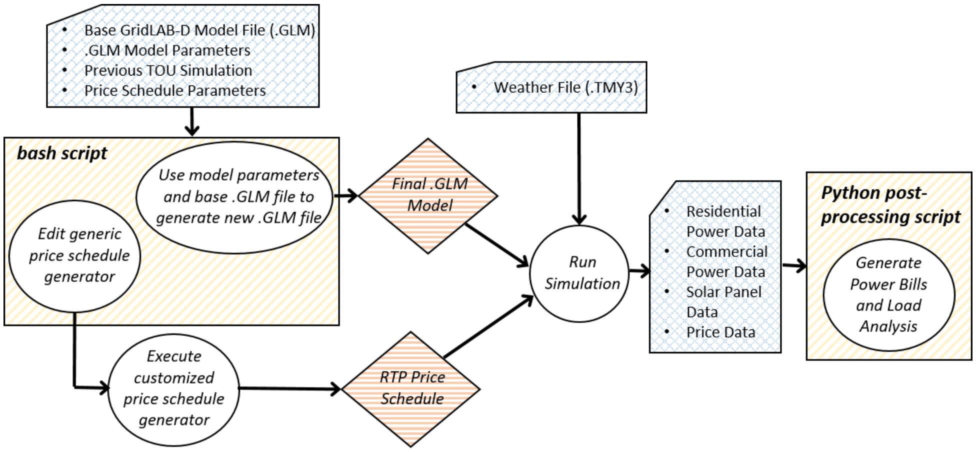

Base GridLAB-D Model File (.GLM)
.GLM Model Parameters
Previous TOU Simulation
Price Schedule Parameters
Weather File (.TMY3)
bash script
Use model parameters and base .GLM file to generate new .GLM file
Edit generic price schedule generator
Final .GLM Model
Run Simulation
Residential Power Data
Commercial Power Data
Solar Panel Data
Price Data
Python post-processing script
Generate Power Bills and Load Analysis
Execute customized price schedule generator
RTP Price Schedule

Figure 2: Simulation Automation Workflow

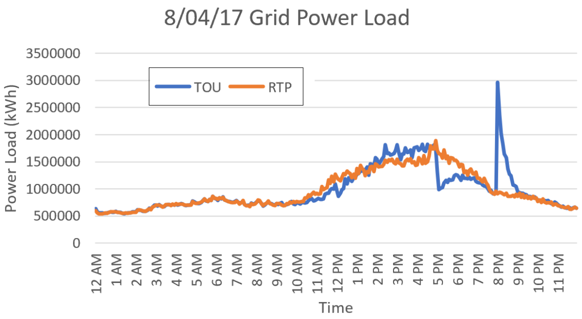


Figure 3: TOU vs RTP Comparison

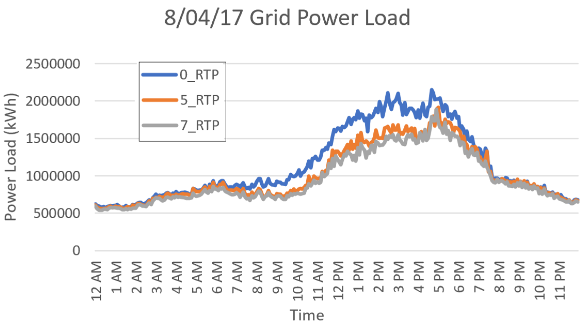


Figure 4: Solar Panel Power Comparison

(2) **Battery storage**: Batteries also have a significant impact on smoothing the duck curve for much of the same reasons that they reduce costs for residents with solar panels. Houses with batteries will use less grid power during peak hours due to the raised prices during those time-frames.

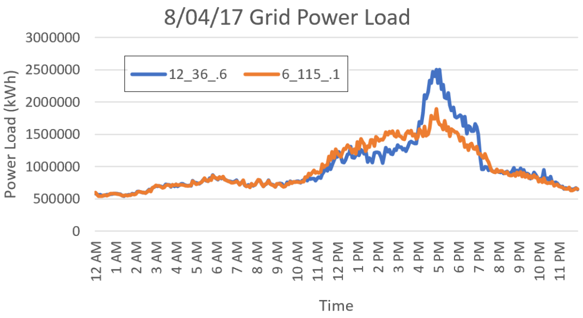


Figure 5: Effect of Tuning RTP Parameters

(3) **Precooling**: The results show that precooling has a minor smoothing effect. Each RTP simulation with precooling has a slightly lower standard deviation and $MaxMin_d$ difference than the same simulation without precooling, but the duck curves are not significantly different.

(4) **Solar panel generation capacity**: Higher power generation capacity of the solar panels reduced grid power demand, as shown in Figure 4, especially during the middle of the day when most of the solar power is produced. Increasing the number of solar panels (50% as opposed to 25%) has a very similar effect on the duck curve as increasing the power generation capacity of the solar panels.

6.3 Effect of Tuning Price Schedules

Through tuning the parameters of the RTP price generator, it was possible to smooth the duck curve significantly. Figure 5 shows a comparison between an RTP simulation with the initial guess for parameters (12_36_.6: x = 12; y =36; and w1 = 0.6) and the set of parameters selected after testing (6_115_.1: x = 6; y = 115; and w1 = 0.1) as described earlier in equations (1), (2), and (3).

Neema, Himanshu; Phillips, Scott; Lee, Dasom; Hess, David; Threet, Zachariah; Roth, Thomas; Nguyen, Cuong. "Transactive Energy and Solarization: Assessing the Potential for Demand Curve Management and Cost Savings." Presented at Design Automation for CPS and IoT (DESTION 2021), Nashville, TN, US. May 18, 2021 - May 18, 2021.

7 CONCLUSIONS & FUTURE WORK

The study demonstrates the benefits of transactive energy with real-time pricing in the context of a local electricity network with distributed solar energy and virtual power plant. The systematic approach to designing RTP pricing shows how the daily demand curve can be made smoother under specifiable conditions. However, the effect on customer prices varies with respect to the use of solar panels, and more research is needed to understand how customers can benefit from the arrangement. The study examined diverse configurations, but many more are possible, and future research could examine additional effects on the two goals of a smoother demand curve and customer pricing benefits.

Further research could also improve upon RTP effects demonstrated by this paper by tuning the price schedule parameters for each simulation. More specific RTP parameters for each model could smooth the duck curve even better than shown in the presented results. With respect to policy implications, the study indicates that utilities and regulators should continue to engage in both simulation experiments and real-world experiments to understand better the effects of transactive energy with real-time pricing. These experiments should include conditions that continue to explore the combinations of battery storage, levels of solar power generation capacity and penetration, and pricing strategy.

8 ACKNOWLEDGMENTS

This work at Vanderbilt University is supported in part by the Partnership for International Research and Education (PIRE) program and the Cyber-Physical Systems (CPS) program of NSF under award #1743772, and by National Institute of Standards and Technology (NIST) under award 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.

REFERENCES

[1] 2018. EIA 2018: Average Monthly Bill-Residential, URL: https://www.eia.gov/electricity/sales_revenue_price/pdf/table5_a.pdf.

[2] 2021. Controller (Transactive Controls). http://gridlab-d.shoutwiki.com/wiki/Transactive_controls

[3] 2021. Weather Data by Location. Retrieved February 16, 2021 from https://energyplus.net/weather-location/north_and_central_america_wmo_region_4/USA/CA/USA_CA_Sacramento.Exec.AP.724830_TMY

[4] Mudathir Funsho Akorede, Hashim Hizam, and Edris Pouresmaeil. 2010. Distributed energy resources and benefits to the environment. *Renewable and sustainable energy reviews* 14 (2010), 724–734.

[5] Chuck Booten, Joseph Robertson, Dane Christensen, Mike Heaney, David Brown, Paul Norton, and Chris Smith. 2017. *Residential Indoor Temperature Study.* Technical Report. National Renewable Energy Laboratory, Golden, CO (United States).

[6] Georgios Chalkiadakis, Valentin Robu, Ramachandra Kota, Alex Rogers, and Nick Jennings. 2011. Cooperatives of distributed energy resources for efficient virtual power plants. *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems – Innovative Applications Track (AAMAS 2011),* (2011), 787–794.

[7] D. P. Chassin, K. Schneider, and C. Gerkensmeyer. 2008. GridLAB-D: An open-source power systems modeling and simulation environment. In *2008 IEEE/PES Transmission and Distribution Conference and Exposition.* 1–5. https://doi.org/10.1109/TDC.2008.4517260

[8] Sijie Chen and Chen-Ching Liu. 2017. From demand response to transactive energy: state of the art. *Journal of Modern Power Systems and Clean Energy* 5, 1 (2017), 10–19.

[9] Poria Hasanpor Divshali, Bong Jun Choi, and Hao Liang. 2017. Multi-agent transactive energy management system considering high levels of renewable energy source and electric vehicles. *IET Generation, Transmission & Distribution* 11, 15 (2017), 3713–3721.

[10] Ahmad Faruqui, Ryan Hledik, Sam Newell, and Hannes Pfeifenberger. 2007. The power of 5 percent. *The Electricity Journal* 20, 8 (2007), 68–77.

[11] D.J. Hammerstrom, R. Ambrosio, J. Brous, T.A. Carlon, D.P. Chassin, J.G. DeSteese, R.T. Guttromson, G.R. Horst, Jarvegre O.M., R. Kajfasz, S. Katipmula, L. Kiesling, N.T. Le, P. Michie, T.V. Oliver, R.G. Pratt, S.E. Thompson, and M. Yao. 2007. *Pacific Northwest GridWise™ Testbed Demonstration Projects.* Technical Report. Pacific Northwest National Laboratory, Richland, WA (United States).

[12] David Hess and Dasom Lee. 2020. Energy decentralization in California and New York: Conflicts in the politics of shared solar and community choice. *Renewable and Sustainable Energy Reviews* 121 (2020), 109716.

[13] J.S. Homer, S.R. Bender, and M.R. Weimar. 2016. *Energy Policy Case Study – California: Renewable and Distributed Energy Resources.* Technical Report. Pacific Northwest National laboratory, Alexandria, VA (United States).

[14] Battelle Memorial Institute. 2015. *Pacific Northwest Smart Grid Demonstration Project Technology Performance Report. Volume 1: Technology Performance.* Technical Report. Battelle Memorial Institute Pacific Northwest Division, Richland, WA (United States).

[15] S. Katipamula, B. Akyol, A. Makhmalbaf, C.D. Corbin, C. Allwardt, V.V. Mendon, J. Haack, B. Carpenter, H. Ngo, H. Hao, S. Huang, S. Somasundaram, W. Kim, G. Liu, Underhill R.M., D.J. Hostick, R.G. Lutes, and M. Zhao. 2017. *Transactive Campus Energy Systems.* Technical Report. Pacific Northwest National laboratory, Alexandria, VA (United States).

[16] Dasom Lee and David Hess. Forthcoming. Data Privacy and Residential Smart Meters: Comparative Analysis and Harmonization Potential. *Utilities Policy* (Forthcoming).

[17] Dasom Lee, David Hess, and Himanshu Neema. 2020. The challenges of implementing transactive energy: A comparative analysis of experimental projects. *The Electricity Journal* 33 (2020), 106895.

[18] Fernando Lezama, Joao Soares, Pablo Hernandez-Leal, Michael Kaisers, Tiago Pinto, and Zita Vale. 2018. Local energy markets: Paving the path toward fully transactive energy systems. *IEEE Transactions on Power Systems* 34, 5 (2018), 4081–4088.

[19] Seyyed Mostafa Nosratabadi, Rahmat-Allah Hooshmand, and Eskandar Gholipourn. 2017. A comprehensive review on microgrid and virtual power plant concepts employed for distributed energy resources scheduling in power systems. *Renewable and Sustainable Energy Reviews* 16 (2017), 641–363.

[20] Eric O'Shaughnessy, Jenny Heeter, Julien Gattaciecca, Jenny Sauer, Kelly Trumbull, and Emily Chen. 2019. Empowered communities: The rise of community choice aggregation in the United States. *Energy Policy* 132 (2019), 1110–1119.

[21] Eric J OShaughnessy, Jenny S Heeter, Julien Gattaciecca, Jennifer Sauer, Kelly Trumbull, and Emily I Chen. 2019. *Community Choice Aggregation: Challenges, Opportunities, and Impacts on Renewable Energy Markets.* Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).

[22] Phillip N Price, Mary Ann Piette, Jessica Granderson, and John Elliott. 2014. Automated Measurement and Verification of Transactive Energy Systems, Load Shape Analysis, and Consumer Engagement. (2014).

[23] KP Schneider, DW Engel, Y Chen, SE Thompson, DP Chassin, and RG Pratt. 2008. Distributed Energy Resources, Virtual Power Plants, and the Smart Grid.

[24] Reza Sharifi, Amjad Anvari-Moghaddam, S Hamid Fathi, Josep M Guerrero, and Vahid Vahidinasab. 2019. An optimal market-oriented demand response model for price-responsive residential consumers. *Energy Efficiency* 12, 3 (2019), 803–815.

[25] Steven E Widergren, Krishnappa Subbarao, Jason C Fuller, David P Chassin, Abhishek Somani, Maria C Marinovici, and Janelle L Hammerstrom. 2014. *AEP Ohio gridSMART demonstration project real-time pricing demonstration analysis.* Technical Report. Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

[26] Maarten Wolsink. 2020. Framing in Renewable Energy Policies: A Glossary. *Energies* 13 (2020), 2871.

[27] Jeff Zethmayr and David Kolata. 2018. The costs and benefits of real-time pricing: An empirical investigation into consumer bills using hourly energy data and prices. *The Electricity Journal* 31, 2 (2018), 50–57.

NIST SP 1283
August 2022

Qualifying Evaluations from Human Operators: Integrating Sensor Data with Natural Language Logs

Michael P. Brundage[1], Michael Sharp[1], and Radu Pavel[2]

[1] *National Institute of Standards and Technology, Gaithersburg, MD, 20899, United States*
michael.brundage@nist.gov
michael.sharp@nist.gov

[2] *TechSolve Inc., Cincinnati, OH, 45237 United States*
pavel@TechSolve.org

Abstract

Even in the increasingly connected world of smart manufacturing and the Industrial Internet of Things (IIoT), there will always be a need for human operators and evaluations. When creating equipment condition monitoring models and heuristics, the observations from human operators are often difficult to quantify or track. This situation can lead to the observations being underutilized, misunderstood, or ignored completely if autonomous sensors are employed. This work seeks to highlight the untapped potential for augmenting numeric data from sensors and control systems with human input and vice versa, by integrating documented natural language reports with data collection technology in a novel and intuitive way. It is a first-step experiment and seeks to establish a link between human-generated data and sensor-driven information to motivate, justify, and guide future endeavors. This is an exploratory work that utilizes an experimental setup with a limited and controlled accelerated aging setup where human observations were recorded at regular intervals alongside streaming sensor data. The goal is to validate the relationship between observers' natural language, quantified sensed values, and some ground truth knowledge about the state of the tool. We provide recommendations for follow-on work and extensions of the performed analysis as part of a next steps outline.

1. Introduction

Currently, many companies, industries, and organizations take advantage of smart manufacturing concepts to design, define, and promote the next generation of digital manufacturing and enterprise capabilities. Large companies are already at the forefront of the development and deployment of digital technologies that enable connectivity and automation (Jin, Weiss, Siegel, & Lee, 2016). However, small- and medium-sized manufacturers (SMMs) lack the resources that big Original Equipment Manufacturers (OEMs) have to research and implement these new concepts (Software, 2018). Without a proper understanding of how emerging technologies and the integration of these technologies can make them more competitive, small manufacturers tend to delay the change of their traditional ways to new, digitally-integrated strategies.

The changes from the digital revolution in manufacturing are profound and pose a real challenge, but also an opportunity to manufacturers of all sizes. To avoid being left behind, companies need to be proactive and develop strategies to exploit the opportunities of digitalization, improve existing processes, and develop new business models.

One of the most rapidly growing trends associated with digital manufacturing is the use of monitoring and data collection systems. These provide visibility and actionable information with respect to machine utilization, capacity, and overall equipment effectiveness. This, in turn, can inform condition-based or predictive maintenance. An increasing number of OEMs leverage technologies such as these to assess the current and future states of equipment, machine tools, manufacturing cells, supporting subsystems, and even manufacturing processes (Software, 2018).

2. Background and Motivation

Researchers have spent considerable time crafting methods to collect and analyze data coming from industrial equipment (Kunche, Chen, & Pecht, 2012; Li, Verhagen, & Curran, 2018). The majority of the research focus is on the analysis of sensor data to improve maintenance operations. Multiple publications provide techniques and results based on numeric data analysis. Kothamasu, Huang, and VerDuin (2006) reviewed the strategies and techniques of monitoring and predicting machine health that focuses on improving reliability and reducing

Brundage, Michael; Sharp, Michael; Pavel, Radu. "Qualifying Evaluations from Human Operators: Integrating Sensor Data with Natural Language Logs." Presented at 6th European Conference of the Prognostics and Health Management Society 2021, Turin, IT. June 28, 2021 - July 02, 2021.

unscheduled downtime. Djurdjanovic, Lee, and Ni (2003) introduced a toolbox of data-driven algorithms and presented applications in mechanical systems prognostics. Katipamula and Brambley (2005) completed a representative review of the research and practices of fault detection, diagnostics, and prognostics for building systems. Lu, Li, Wu, and Yang (2009) summarized wind turbine condition monitoring and fault diagnosis activities. Venkatasubramanian, Rengaswamy, Yin, and Kavuri (2003) presented a review of quantitative model-based methods for chemical process fault detection and diagnosis. More recent methods leveraging machine learning (ML) and artificial intelligence (AI) techniques continue to heavily rely on numeric data collected from sensors, PLCs, and machine controls.

Analyzing human-generated text from the shop floor offers another promising avenue to improve operations. This area of research is called Technical Language Processing (TLP) and centers around using Natural Language Processing (NLP) methods on technical text data (Brundage, Sexton, Hodkiewicz, Dima, & Lukens, 2021). In particular, using TLP on Maintenance Work Orders (MWOs) has been an area of budding research (Brundage et al., 2021; Ho, 2015; Lukens, Naik, Saetia, & Hu, 2019). Information within MWOs provides a health history of an asset that is rich with quantitative and tacit knowledge within the text. Previously, researchers have analyzed this information to capture key information about maintenance operations (Ho, 2015; Lukens et al., 2019; Sexton, Brundage, Hoffman, & Morris, 2017). Multiple efforts successfully calculated the Mean Time Between Failure (MTBF) by only using MWOs (Ho, 2015; Sexton et al., 2017). Other works have created maintenance Key Performance Indicators (KPIs) from MWOs to help analysts understand the performance of their maintenance operations (Brundage, Morris, Sexton, Moccozet, & Hoffman, 2018).

An underexplored area of research is the merging of these two major data streams: (1) data coming directly from the equipment and (2) the human-generated text data. The human-generated test data (i.e. natural language) can add significant value to the maintenance analysis since observations by humans on the floor (e.g., “the tool is smoking” or “there is a banging noise in the machine”) provide context to the sensor data. Hybrid data can improve predictive maintenance capabilities by providing ground truth to the “state” of the component being monitored. As an example, if an estimate shows that a bearing will fail in 5 days but the technician pulls it from the floor with zero damage, an analyst can update their model based on these observations. Similarly, if the sensor data indicates zero problems with the bearing, but the technician observes heavy smoke and noise, this can also improve the predictive model.

Hybrid data has significant potential for artificial intelligence (AI) techniques that can achieve improved accuracy and classification based on information captured through natural language. In addition, text data can complement numeric data by providing information about subsystems that may be outside the reach or sensitivity of the sensor-based systems (e.g., a tube that disconnects every few hours of equipment utilization, or a filter that needs to be changed so the surface of a component does not get contaminated).

This paper aims to illustrate the importance of this hybrid dataset to improve maintenance operations. First, we describe an experimental setup to generate real-world data that combines both text-based data via human observations and sensor data to investigate tool wear. This experiment can be run within any laboratory testbed or manufacturing environment. Second, we provide methods on how to merge this data and prepare it for analysis. This leads to initial insights about analyzing the data and how it can be used to improve maintenance operations. Lastly, we discuss improvements to the experiment and future extensions to this work.

3. TESTBED SETUP

This experimental work uses TechSolve’s M. Eugene Merchant Technology Development Center machining lab to develop a dataset as close to live manufacturing environments as possible. The facility has modern Computer Numerical Control (CNC) machines, a full array of measuring and analysis equipment (including force dynamometers, profilometers, Coordinate-measuring machine (CMM) equipment, microscopes, and inspection devices), and a variety of standard shop machines and equipment. The selected testbed consists of an instrumented machine tool. The experiments use the tool to run cutting tests under the close observation of experienced machine operators. The primary goal of the setup was to rapidly degrade a series of cutting tools under increased workloads while recording both instrument readings and periodic human observations via free-form text. This setup allows establishing relations between the human-generated and sensor-driven data. Periodic direct measurements of the tool-wear were also taken and are used as a “ground truth” basis for the health of the tool-piece in the results section of this paper.

The experimental setup included the following elements:

- Machine: Milltronics HMC35, instrumented with sensors, data acquisition system and tool condition monitoring system
- Cutting tools: Carbide end mill with 4 flutes, 0.5” diameter and 1” length of cutting zone
- Metalworking fluid: Water-based (Trimsol 206)
- Workpiece: 4140 steel block of 6” x 4” x 4”
- Fixturing: The workpiece, clamped into a vise with 6” long jaws

The data collection system enabled simultaneous acquisition of machine control data and data from the added sensors. The following data was collected from the Fanuc 0i Controller: date and time; the axes positions in absolute, machine, and relative coordinates; distance to go on each axis; actual spindle speed; actual feed rate; spindle load; spindle motor speed; exes loads; servo delays; and the acceleration or deceleration delays. The added sensors included:

- A three phase hall effect sensor, monitoring the power drawn by the motor of the spindle with an analog output of 0 to 10VDC corresponding to 0HP to maximum HP
- Uniaxial accelerometers ((Integrated Circuit-Piezoelectric (ICP) type), placed on the housing of one of the ball-screw bearings of each axis
- A tri-axial accelerometer on the spindle housing, an Integrated Electronics Piezo-Electric (IEPE) sensor with a higher sensitivity comparing to the accelerometers on the feed axes
- J-type thermocouples on each axis and each axis motor, on the spindle, and in the metalworking fluid tank

The machining center has other sensors and instrumentation, which were not used for this experiment.

4. EXPERIMENTAL DESIGN

Using the setup described above, the experiment design focused on linking sensor values with simultaneous human observations. The experiment focused on the degradation of the cutting tools rather than the degradation of the machine itself, allowing repeatable and timely observations. The approach included the following steps:

1. Create a long cut machining program of at least 1.5 hours.
2. Modify the program such that some of the cuts exhibit chatter or higher than normal vibration – to create controlled process failures.
3. While the CNC program is running, collect data from sensors and controllers and have a technician, other than the one that created the machining program, come to observe the test periodically (e.g., 15 minutes) and take notes relative to the status of the cut, tool and machine (similar to creating maintenance logs).

To create the cutting program, TechSolve engineers identified the test conditions for this experiment through a series of exploratory cutting tests. These tests validated the physical combination of material, tooling, and cutting parameters, and helped establish the intervals for collecting human-generated information during the experiment. The results defined a CNC program that would enable a cut of approximately 3 hours (continuous cutting) including regions with chatter or increased vibration for the tool to simulate real-world observable events. The exploratory cutting tests also established a criterion for the end of tool life. The engineers observe the exploratory tests and use these observations to measure tool wear after machining one complete surface (one layer). A cutting test was considered complete after removing 6 layers of material.

The engineers use climb milling to machine the 6” x 4” surface of the workpiece (the steel block) in a transversal direction. A spindle tap test determined the chatter lobe diagram and the stability zone for the tool-spindle assembly. Based on this information, the team identified cutting conditions that would generate chatter. The CNC program’s design induced chatter to introduce abnormal cutting conditions in the experiment. Chatter was achieved by increasing the radial depth of cut and slightly increasing the feed rate. The team used a limited number of test scenarios to validate chatter generation and to observe the effects on the cutting tool and machined surface. Eventually, the team selected the cutting conditions presented in Table 1 for both the normal and abnormal (chatter) cuts.

The tests were planned and conducted as follows. One technician (Technician A) participated in the development of the test and of the CNC program, which included randomly inserted abnormal (chatter) cutting parameters. Another technician (Technician B) conducted observations on the cutting process at periodic intervals.

Each cutting test followed the procedure below:

1. The machine setup was prepared and the workpiece was clamped in the vise by Technician A.
2. Technician A took pictures of the fresh cutting tool prior to starting the test; both the end and lateral surfaces engaging the workpiece have been photographed.
3. Technician A started the CNC program, while Technician B observed and took notes of the process condition.
4. The machine ran linear climb milling cuts according to the CNC program created by Technician A.
5. While Technician B was free to move away from the machine, they were instructed to make text entries about the operation every 15 minutes. The technician was unaware prior to any given cutting pass if it was being performed with the chatter parameters. At 30 minute intervals, once a layer was removed from the steel block, Technician B had the option to pause the program and observe the tool and workpiece condition. A table was created in Excel to record various characteristics of the process. The idea was to emulate a data collection log similar to what is applied for MWOs. Table 2 lists a selection of the recorded characteristics.
6. A test was considered complete after 6 layers were removed from the steel block. A layer consisted in a volume of 6” x 4” x 0.5”, removed through the milling process with the 0.5” axial depth of cut.

The experiment automatically collected data from the sensors and machine tool control for each pass. Technician B was the main technician for observing the tests and collecting notes.

Table 1. Cutting Parameters

Parameter	**Normal Cut (Stable)**	**Abnormal Cut (Chatter)**
Axial Depth of Cut	0.5 in	0.5 in
Radial Depth of Cut	0.04 in	0.12 in
Tool Diameter	0.5 in	0.5 in
Cutting Speed	435 sfm	435 sfm
# of Teeth	4	4
RPM	3323 rpm	3323 rpm
Feed per Tooth	0.0018 in/tooth	0.002 in/tooth
Feed Rate	24.00 in/min	26.59 in/min
Feed Stroke	4.75 in	4.75 in

Table 2. Human Recorded Process Characteristics

Recorded Characteristic	**Description**
Test No. and Workpiece No.	Indication Value [1 - 6]
Various Test and Observation Times	Time Records
Test Status	Short Descriptor
Layer No.	Indication Value [1st / 2nd / etc]
Tool Gauge length (in)	Value
Type of observation	Short Descriptor
Process condition	Short Descriptor [Normal / Abnormal / other]
Tool condition	Short Descriptor
Tool flank wear - END (in)	Value [0 - 0.005 in]
Tool flank wear - LATERAL (in)	Value [0 - 0.005 in]
Tool wear - RAKE FACE (in)	Value [0 - 0.005 in]
Surface finish - face (μ in)	Value [$\sim$15 - 50 Ra]
Surface finish -lateral (μ in)	Value [$\sim$15 - 25 Ra]
Did you hear chatter?	Yes / No
Anything out of ordinary?	Yes / No
Out of ordinary description	Short Description
General Notes/Comments	Free Form Text

However, if Technician B was unavailable due to other tasks, Technician A would take notes. In general, the technicians were instructed to write down observations and, if needed, stop the cycle to observe what happened if they heard sounds or observed unusual behavior of the cut. Irrespective of observing irregularities or not, the test was stopped at the end of each machined layer (approximately 30 minutes) to allow Technician B to measure tool wear. Either technician would then start the process where it was left off, and the process of collecting numeric data and periodical observations continued.

5. RESULTS

We aim to establish a quantitative link between the recorded human observations and the sensed values of the test pieces. Establishing this link requires translating the human observations into some ordinal scaled value. Although we expect this to eventually be automated, we accomplished translation by having independent experts read the human-generated text. After the tests were completed, the experts read the entries to infer a level of damage between None to Imminent Failure. We presented text entries in the *Recorded Characteristics* table (Table 2) to 7 independent experts to interpret. They were instructed to disregard any measured wear or finish quality values and focus exclusively on the text entries. Figure 1 shows an example of the interpreted damage from the 7 experts. We use the average value of the expert interpretations to calculate the quantitative link between the sensor values, "ground truth", and the human observations.

The manually-measured values of the tool wear are used as the "ground truth" level of degradation on each machine tool. Figure 2 shows that there is a strong correlation between human-interpreted damage level and the measured wear of the tool. The average correlation between the interpreted damage and the measured wear across all tests is 0.74, with a maximum of 0.97 and a minimum of 0.42. This includes all statistically significant correlation coefficients (p ¡ 0.05) for all of the three measured wear features: rake face, flank lateral, and flank end wear. Not all values were measured at all points. When there were not enough recordings to establish statistical significance, we omitted the correlation coefficient.

The next step focuses on quantifying the relationship between sensed values and human interpretations. Due to the asynchronous nature of the two sets of data, we employed a method for determining asynchronous correlation. This simple process uses dual signal interpolation to find the estimated overlap values for each data stream, then concatenates those values to determine a correlation coefficient. We present the breakdown

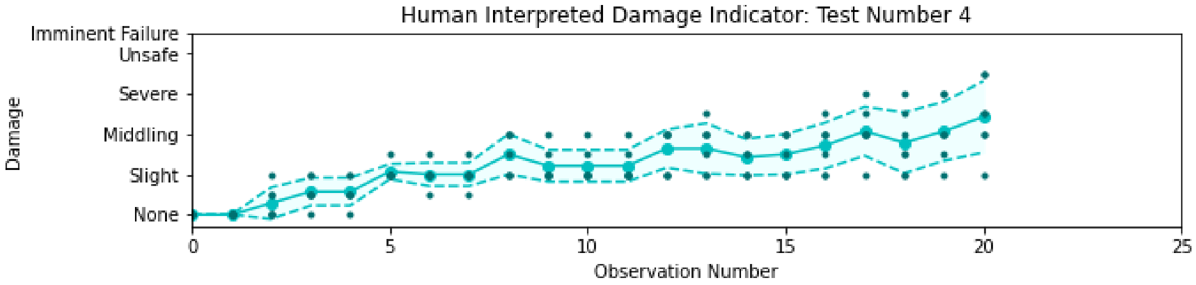


Figure 1. Results of Human Interpreted Damage for Test Piece 4

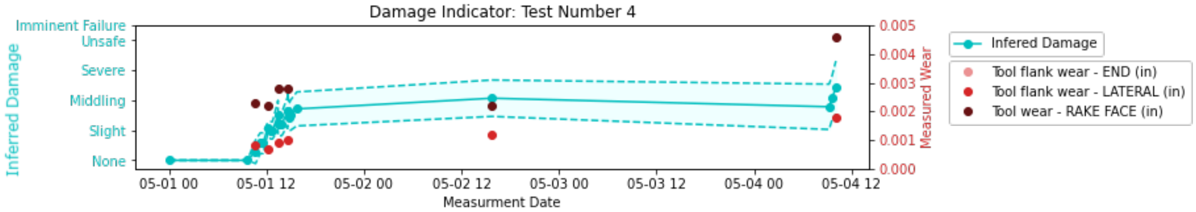


Figure 2. Relationship Between Human Interpreted Values and Ground Truth Wear
(END: Corr = nan, P = nan) (LATERAL: Corr = 0.80, P = 0.00) (RAKE FACE: Corr = 0.66, P = 0.00)

of that process as pseudo code below.

Algorithm 1: Pseudocode

```
With S1 = Signal 1 Values, S1t = Signal 1 Locations
With S2 = Signal 2 Values, S2t = Signal 2 Locations
# Find Points of Overlap
OL1 = t < max(S2t) & t >= min(S2t) for t in S1t;
OL2 = t < max(S1t) & t >= min(S2t) for t in S2t;
# Interpolate Sig X at Sig Y locations
if len(S1)>1 then
    F1 = interpolate(S1t,S1, kind = interpkind);
    S1est = F1(S2t(OL2));
else
    S1est = S1[0] for x in S2t(OL2);
end
if len(S2)>1 then
    F2 = interpolate(S2t,S2, kind = interpkind);
    S2est = F2(S1t(OL1));
else
    S1est = S2[0] for x in S1t(OL1);
end
# Concatenate and Calculate Correlation
S1cat = concatenate((S1(OL1),S1est),axis = 0) ;
S2cat = concatenate((S2(OL2),S2est),axis = 0) ;
return Correlation(S1cat,S2cat)
```

The goal of this work was not to establish an optimal degradation inference method from sensor signals, but to show that correlating the human observations with information derived from sensed values is not only possible, but may yield more robust methods for decision support than either could provide alone. Raw sensor data rarely is used in live settings for decision support, and so the decision was made to use a simplistic form of information extraction for the recorded sensors. This method compressed standard-sized windows of the raw signals into one root mean square value per window. This step both acts as a pseudo information extraction algorithm and makes correlating the signals more visually appealing with less computational demand due to the reduction of raw values. Future work may focus on more robust information extraction methods to connect the sensor values to the ground truth degradation of the tool wear.

Because of this simplistic data compression/ information extraction method, we expect a noticeable level of disconnect between the recorded values and the ground truth. However, the authors feel there is sufficient correlation in enough of the signals to establish that concurrent information exists.

This work focused on signals whose activity was expected to directly correspond to the wear of the tool. This included the power signal, the tri-axial accelerometer on the spindle, three accelerometers on the feed axes, and a suite of temperature sensors described in the previous section. The figures below show an example of the processed signals and their correlations between other sensor signals.

Inspecting Figures 1, 2, 3, we can visually identify a pattern of increasing values over time occurring in the human observations, the “ground truth”, and the sensed values. Asynchronous correlation between the respective values confirms this visual cue.

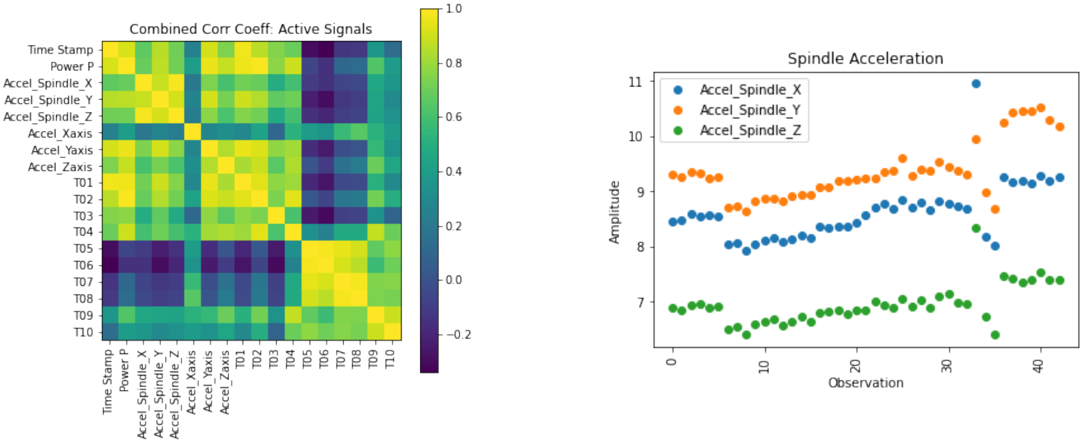


Figure 3. Relationships of Sensed Signals

Figure 4 shows the relationship for both the power consumption and the spindle acceleration to the human observations. These exhibit strong asynchronous correlations of 0.92 and 0.91, respectively.The average correlation between each group of signals and the human assessment can be found in Table 3.

Looking at this table, it is very important to note that most of the negative or low correlation values are due to the imprecision of the information extraction method applied to the sensors and the disregard for confounding factors. This is especially pertinent in the cases of the temperature signals where long delays between starts and stops can have a profound effect on the pattern of the signal. See Figure 5 as an example of this.

Notice that even in these examples that the signals visually trend with the human assessment inferred damage values. These trends, in turn, follow the ground truth pattern of the measured wear values and establish a clear link between the wear of a tool and the human assessment of that tool in a live working environment.

The visual confirmation, the calculated correlations, and the intuitive understanding of how humans process information provides ample evidence to support the idea that asynchronous human input should be used to supplement sensor readings in a live environment. This work justifies the exploration of deeper topics relating to the combination of heterogeneous data sources, both from human and mechanical sources. The value of using human agents as supplemental sensors in areas that are under-sensed or hard to evaluate makes intuitive sense. However, there is also value to adding that information to a well-sensed environment or asset. Humans are natural pattern identifiers and can add valuable insight or tacit information that a limited sensor set might not be able to. Future work will explore methods for incorporating this information and quantifying the value return of its use.

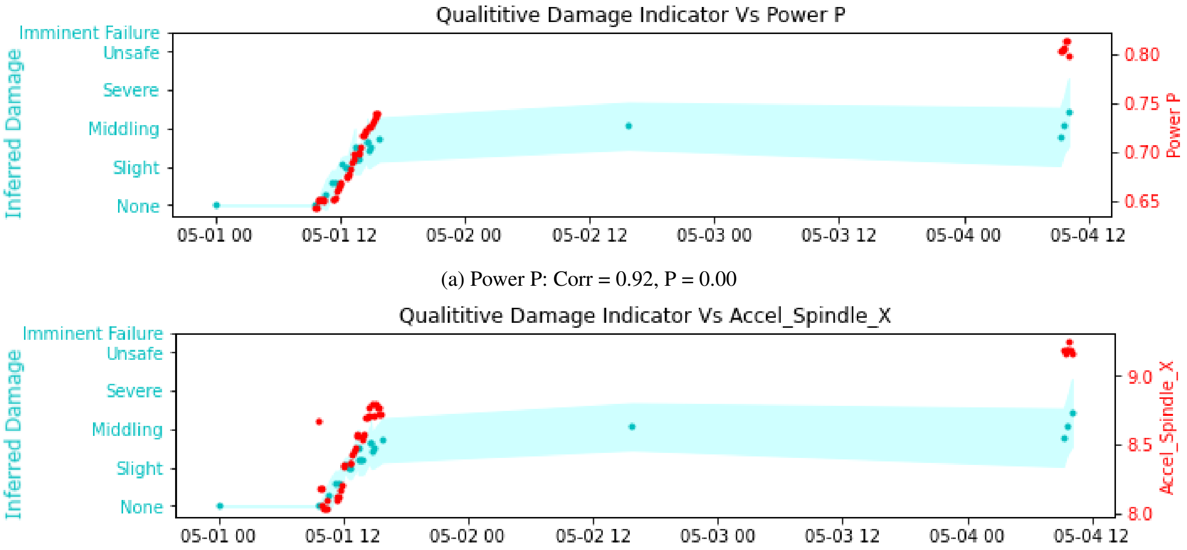
6. DISCUSSION

This experimental setup and initial results provided some significant observations and potential avenues for further improvements.

Adding More Human Observers The major focal point for creating this dataset is the linking of the human observations to mechanically-sensed values. Of the two, the highest levels of variations will arise from the human observations, and thus, a greater number of human observations will allow for better relational development with the more rigorously sensed data. Although a high level of human observations or redundant observations may not be expected in a real-world scenario, adding more human observers would greatly enhance the development and validation of methods and technologies to best use this information in a real-world setting.

The current setup involved one human observer at any given observation of a machine. Future experiments can explore having multiple humans to simulate a real maintenance operation. The ideal dataset would have multiple people independently evaluating and recording regular observations of each machine through time. This would allow for developing better uncertainty bounds and expected variations between human agents.

Future investigations might also seek to find the saturation point for inserting human observations. Intuitively, humans will not effectively notice subtle changes over time if the change is gradual enough; the same is true for some computer-based monitoring systems. However, allowing a human agent to step away from the system, then check back on it after some interval may circumvent this problem. Identifying the interval of greatest return



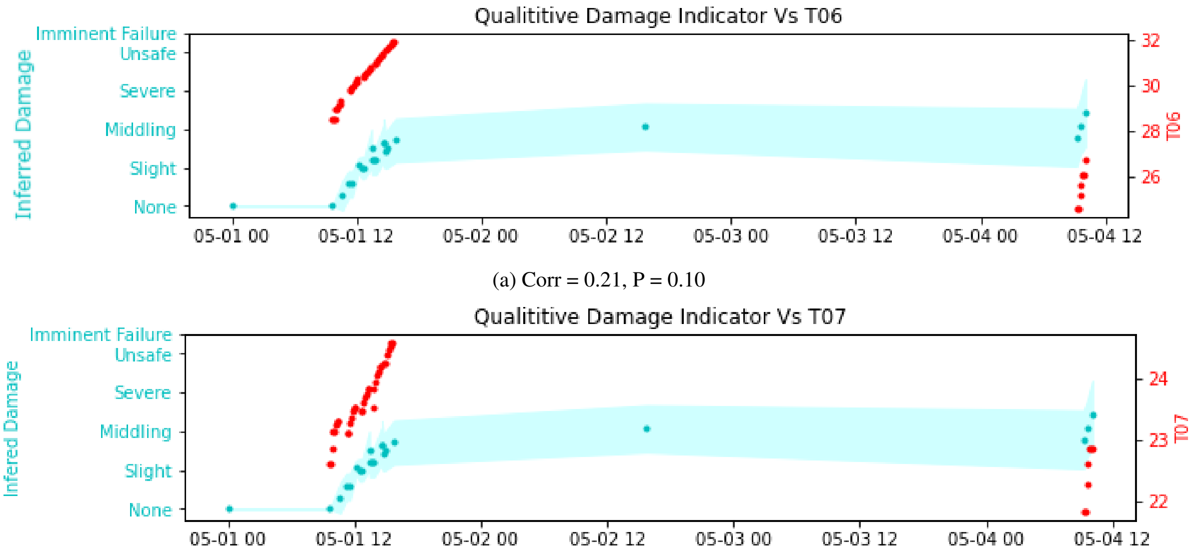
(a) Power P: Corr = 0.92, P = 0.00

(b) Accel Spindle X: Corr = 0.91, P = 0.00

Figure 4. Relationship Between Sensed Signals and Human Assessment

Table 3. Asynchronous Correlations Between Sensors and Human Assessment

	Correlation to Human Damage Assessment		
Sensor Group	**Average**	**Max**	**Min**
Power	0.44	0.92	-0.69
Spindle Accelerometers	0.66	0.91	0.22
Machine Accelerometers	0.64	0.85	0.39
Temperature	0.11	0.95	-0.79



(a) Corr = 0.21, P = 0.10

(b) Corr = 0.09, P = 0.50

Figure 5. Example of Temperature Shift Caused by Delayed Restart

from human observation could help to schedule regular 'walk-through' style checkups on various processes and machines. This process would be especially useful for assets that are not fully equipped with mechanical sensors.

Incorporating NLP/TLP Humans are best suited to express their observations through free-form text. Although Likert scales and similar can be useful in analyses, they are prone to inconsistency and lack much of the contextual information that free-form text can provide Hodge and Gillespie (2003). Additionally, much of the human-collected information already available in industrial settings (e.g., maintenance work orders) does not intrinsically contain this type of structured information. This motivates and necessitates a focus on natural language processing (NLP) or technical language processing (TLP) as a means to automate and capture a more full scope of any human observation.

By its very nature, free-form text is somewhat inconsistent, and as such, difficult to definitively confirm ground truth. To help circumvent this, this work created a post-collection Likert scale value for each entry from a human observer. These were made by aggregating the responses of multiple experts, but fundamentally may or may not have captured the original observer's intent. An ideal test setup would prompt the human observers to provide some Likert-style value as well as the free-form text. This setup, coupled with the post-collection interpretations, could be used to train and validate any developed NLP/TLP models as well as provide more information about the expected uncertainty within the data.

Developing state-of-the-art language processing models could enable deeper and more rich use of currently available data. These would eventually preclude the need for Likert-style assessment from the human operators. These advances can allow observers to express their assessment more efficiently. In turn, this procedure increases the chances that the observer will provide useful information.

Focus on Temporal Asynchronicity Facilities often sense and record values at unaligned intervals. Adding in irregular and inconsistent observations and recordings from human agents creates a strong need to address methods for merging asynchronous data streams. This work showed visual alignment, as well as one method for interpolating expected values to create alignment. However, these are not the only, nor the best possible methods.

Addressing asynchronicity can be done in a multitude of ways that will largely be dictated by the desired results and the preferred models. Although it would take very little effort to find a solution for a given application, future work may want to focus on finding the optimal methods to aggregate, incorporate, merge, or align heterogeneous data sources that are largely asynchronous.

Experimental Scope The scope of any future experiments should build upon those that have come previously. The current experiment focused on a single type of machine tool, specifically a CNC milling center. Conversely, a future experiment could focus on a series of machines to provide scenarios similar to a full manufacturing line. Such a setup would allow researchers to not only understand process parameters but manufacturing system dynamics as well.

Future experiments should first recreate the scope of the experiment described in this work with a larger number of test units. This work's experiment involved a single type of failure on a single type of asset. Next, experiments should progress to multiple types of failures on a single asset. Later experiments could address multiple failure types amongst multiple assets. We recommend this progression of experiments because it allows for steady validation of developed tools while retaining focus on the human-supplied portions of the experimental data.

In any experiment, a minimum number of assets should be subjected to trials to develop statistical significance and allow for variation across the human observations. Although needs may vary depending on specific setups, the authors suggest a minimum of 50-100 entries confirmed by some 'ground truth' as a starting point. NLP or TLP tools may require hundreds or even thousands of entries Brundage et al. (2021).

Real-World Data Concerns Although controlled testing and environments greatly ease the process of defining and developing tools and technologies for incorporating human-derived information, the perspective that these will ultimately be used in a live industrial setting should not be lost. Whenever possible, steps should be taken to ensure that the types and formats of data ultimately reflect those that could or would be acquired in an industrial setting.

Unfortunately, obtaining real-world data, particularly industrial data, can be difficult due to proprietary restrictions and fear of losing competitive advantage. Whenever possible, providing real-world datasets can help researchers advance their analysis methods by verifying their tools and test datasets against real industrial data. Making reference datasets - both laboratory and real-world - available to the general public can accelerate the development of applicable tools, best practices, and standards.

7. CONCLUSIONS

This paper discusses the need and a methodology to create a dataset with both sensor-based and human-generated data.

Our initial analysis illustrates the value of capturing this information within the scope of maintenance operations. Our results show a strong correlation between a human interpretation of the system, ground truth measurements, and hard sensor values captured during the experiment.

This work provides the initial motivations and justifications for further developing rigorous methods to utilize human-derived data in the traditionally-incompatible environment of sensor-driven technologies. We provide discussions on successes and challenges faced during this experiment, along with a loose guide on improvements for future work.

Skilled humans will always be one of the most accurate tools for assessing the broadest intake of direct and indirect information about a system. Sensing equipment can provide more consistent and objective precision than any single human. Each is well suited to provide incredibly useful information for their respective areas of excellence. The challenge we highlight and begin to address is the development of datasets that show this. Datasets facilitate the development of tools and technologies that capture and capitalize on both types of valuable information. The future of industry lies at the intersection of humans and technology.

ACKNOWLEDGEMENT

The authors would like to thank Timothy Sprock and Moneer Helu from NIST for their guidance in crafting the experiment.

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

Brundage, M. P., Morris, K., Sexton, T., Moccozet, S., & Hoffman, M. (2018). Developing maintenance key performance indicators from maintenance work order data. In *International manufacturing science and engineering conference* (Vol. 51371, p. V003T02A027).

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, *27*, 42–46.

Djurdjanovic, D., Lee, J., & Ni, J. (2003). Watchdog agent—an infotronics-based prognostics approach for product performance degradation assessment and prediction. *Advanced Engineering Informatics*, *17*(3-4), 109–125.

Ho, M. (2015). *A shared reliability database for mobile mining equipment* (Unpublished doctoral dissertation). University of Western Australia.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to likert scales. *Social Work Research*, *27*(1), 45–55.

Jin, X., Weiss, B. A., Siegel, D., & Lee, J. (2016). Present status and future growth of advanced maintenance technology and strategy in us manufacturing. *International journal of prognostics and health management*, *7*(Spec Iss on Smart Manufacturing PHM).

Katipamula, S., & Brambley, M. R. (2005). Methods for fault detection, diagnostics, and prognostics for building systems—a review, part i. *Hvac&R Research*, *11*(1), 3–25.

Kothamasu, R., Huang, S. H., & VerDuin, W. H. (2006). System health monitoring and prognostics—a review of current paradigms and practices. *The International Journal of Advanced Manufacturing Technology*, *28*(9-10), 1012–1024.

Kunche, S., Chen, C., & Pecht, M. (2012). A review of phm system's architectural frameworks. In *The 54th meeting of the society for machinery failure prevention technology, dayton, oh.*

Li, R., Verhagen, W. J., & Curran, R. (2018). A functional architecture of prognostics and health management using a systems engineering approach. In *Proc. eur. conf. phm soc* (pp. 1–10).

Lu, B., Li, Y., Wu, X., & Yang, Z. (2009). A review of recent advances in wind turbine condition monitoring and fault diagnosis. In *2009 ieee power electronics and machines in wind applications* (pp. 1–7).

Lukens, S., Naik, M., Saetia, K., & Hu, X. (2019). Best practices framework for improving maintenance data quality to enable asset performance analytics. In *Annual conference of the phm society* (Vol. 11).

Sexton, T., Brundage, M. P., Hoffman, M., & Morris, K. C. (2017). Hybrid datafication of maintenance logs from ai-assisted human tags. In *2017 ieee international conference on big data (big data)* (pp. 1769–1777).

Software, S. P. (2018). "the connection between production monitoring and oee" [Computer software manual]. (Available at `https://www.engineering.com/story/40841`. Accessed 04-05-21)

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & chemical engineering*, *27*(3), 293–311.

OpenASR20: An Open Challenge for Automatic Speech Recognition of Conversational Telephone Speech in Low-Resource Languages

Kay Peterson[1], *Audrey Tong*[1], *Yan Yu*[2]

[1]National Institute of Standards and Technology, USA
[2]Dakota Consulting Inc., USA
kay.peterson@nist.gov, audrey.tong@nist.gov, yan.yu@nist.gov

Abstract

In 2020, the National Institute of Standards and Technology (NIST), in cooperation with the Intelligence Advanced Research Project Activity (IARPA), conducted an open challenge on automatic speech recognition (ASR) technology for low-resource languages on a challenging data type - conversational telephone speech. The OpenASR20 Challenge was offered for ten low-resource languages - Amharic, Cantonese, Guarani, Javanese, Kurmanji Kurdish, Mongolian, Pashto, Somali, Tamil, and Vietnamese. A total of nine teams from five countries fully participated, and 128 valid submissions were scored. This paper gives an overview of the challenge setup and procedures, as well as a summary of the results. The results show overall high word error rate (WER), with the best results on a severely constrained training data condition ranging from 0.4 to 0.65, depending on the language. ASR with such limited resources remains a challenging problem. Providing a computing platform may be a way to level the playing field and encourage wider participation in challenges like OpenASR.

Index Terms: automatic speech recognition, evaluation, low-resource language, conversational telephone speech, IARPA MATERIAL, Amharic, Cantonese, Guarani, Javanese, Kurmanji Kurdish, Mongolian, Pashto, Somali, Tamil, Vietnamese

1. Introduction

The performance of ASR technologies has been under investigation for decades. NIST started conducting benchmark ASR tests in the 1980s, beginning with English read speech in limited domains. In collaboration with the Spoken Language Program of Defense Advanced Research Projects Agency (DARPA), a series of Large Vocabulary Continuous Speech Recognition (LVCSR) tests took place in the 1990s, over time adding more data and data genres such as broadcast news and conversational speech, as well as other high-resource languages such as Arabic and Spanish. Overviews of the progression of these tests can be found in [1], [2]. The DARPA Effective, Affordable, Reusable Speech-to-Text (EARS) Program that ran from 2002 to 2004 also marked the beginning of the Rich Transcription (RT) evaluation series that ran from 2002 to 2009[3]. From 2006 to 2011, the DARPA Global Autonomous Language Extraction (GALE) program, while not having an ASR main focus, had ASR as a component and continued to further speech recognition evaluation. Performance improved over years of repeated testing, with some languages and genres reaching a similar performance as human transcription of speech.

As with human language technologies in general, ASR presents a bigger and different challenge if less is data available for training and development. There are over 7000 languages spoken in the world; the vast majority of them are considered low-resource. At the same time, improved performance for such languages is becoming increasingly more important, with more widespread use of devices and technologies in more languages for which technologies such as ASR are an integral part of the user experience. The performance of ASR also affects the quality of downstream applications such as machine translation. As human language technologies are maturing, the challenges of low-resource language technologies have come more into focus as a subject of increased research interest in the 2000s, as surveyed e.g. in [4]. The Workshop on Spoken Language Technologies for Under-Resourced Languages, held biannually since 2008 and most recently as the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages (CCURL) in 2020, has been a venue for continued research into low-resource speech technologies including speech recognition.[5] The IARPA Babel program, which ran from 2012 to 2016, tested rapid development of ASR and keyword search technologies for conversational telephone speech in languages with little transcribed data available.[6]

Against this background, the OpenASR Challenge[7] hosted by NIST is an open evaluation challenge created as a spin-off of the IARPA Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program[8], which encompasses more tasks, including cross-language information retrieval, language and domain identification, and summarization. For every year of MATERIAL, NIST supports a simplified, smaller scale evaluation open to all, focusing on a particular technology aspect of MATERIAL. The capabilities tested in the open challenges are expected to ultimately support the MATERIAL task of effective triage and analysis of large volumes of text and audio content in a variety of less-studied languages.

In 2020, the focus was on assessing the state of the art of ASR for low-resource languages. Ten low-resource languages were offered. The task was to perform ASR on speech data in these languages, producing written text output. OpenASR20 was implemented as a track of NIST's Open Speech Analytic Technologies (OpenSAT) evaluation series.[9] The evaluation made use of existing data, thus offering a low-cost option to perform an evaluation on a multitude of languages.

2. Challenge Setup

2.1. Languages

The OpenASR20 Challenge was offered for these ten low-resource languages (shorthand used in results in parentheses): Amharic (AMH), Cantonese (CAN), Guarani (GUA), Javanese (JAV), Kurmanji Kurdish (KUR), Mongolian (MON), Pashto (PAS), Somali (SOM), Tamil (TAM), and Vietnamese (VIE).

Apart from being considered low-resource, several of these

Peterson, Kay; Tong, Audrey N.; Yu, Yan. "OpenASR20: An Open Challenge for Automatic Speech Recognition of Conversational Telephone Speech in Low-Resource Languages." Presented at INTERSPEECH 2021: Special Session on OpenASR20 (Open Automatic Speech Recognition) and Low Resource ASR Development, Brno, CZ. August 30, 2021 - September 03, 2021.

languages exhibit additional challenges that often come with low-resource status, such as inconsistent spelling conventions, underspecified orthographies, potential code-switching, and dialectal variations. Participants could attempt as many of the ten languages as they wished.

2.2. Training Conditions

The OpenASR20 Challenge offered two different training conditions, Constrained (CONSTR) and Unconstrained (UNCONSTR). For any language processed, participants were required to make a submission for the Constrained Training condition. The Unconstrained Training condition was optional, but encouraged.

As the name implies, the *Constrained* training condition limited training data resources to allow better cross-team comparisons. The constraint was quite severe; the only speech data permissible for training under this condition was a 10-hour training set specified by NIST in the language being processed. Additional text data in any language, either from the provided training set or publicly available resources, was permissible for training in the Constrained Training condition. Any such additional text training data had to be specified in sufficient detail in the system description. Under the *Unconstrained* training condition, participants were allowed to use speech data outside of the provided 10-hour training set, as well as additional publicly available speech and text training data from any language. This condition allowed for gauging performance gain from additional training data. Any such additional training data had to be specified in the system description.

Participants were not allowed to hire native speakers for data acquisition, system development, or analysis for either training condition.

2.3. Data

The data used in the challenge consisted of conversational telephone speech between two individuals speaking on a topic of their choosing from a list of suggested topics, for approximately ten minutes. Separate datasets for system training (BUILD), development (DEV), and evaluation (EVAL) were provided for each of the languages. The conversations were provided as separate channels for each speaker. For a few cases, only one side of the conversation was available due to not passing quality control checks. The data were sampled at 8kHz, 44.1kHz, or 48kHz and provided in .sph or .wav format, depending on the language, and were marked for gender, age, dialect, environmental condition, network, and phone model. The BUILD set also included a lexicon and a language specification document. Given the low-resource status of the languages, there are varying degrees of spelling variation and inconsistencies to be found in the data - adding an additional level of challenge, despite the fact that transcriptions were conventionalized in a standardized orthography. The data for most of the languages stem from the IARPA Babel program. More details regarding the audio data can be found in the IARPA Babel Data Specifications for Performers.[10] The Somali data sets stem from the IARPA MATERIAL program. For a larger overview of the corpora used in MATERIAL, including ASR and other data, see [11].

Prior to scoring OpenASR20 submissions, the data was preprocessed with a number of normalization steps. These steps are detailed in section 7.3 of the OpenASR20 Evaluation Plan.[12]

Table 1 lists the permissible BUILD (CONSTR and UNCONSTR), DEV, and EVAL data resources by modality, audio vs. text. The same limits held for all ten languages. The audio durations listed refer to conversations, not total length of audio files (which were separate for each speaker). Table 2 lists the number of audio files and the approximate number of words for each of the ten languages' BUILD, DEV, and EVAL data sets.

Table 1: *BUILD, DEV, and EVAL data resources. Hours of audio indicate duration of conversation, not total duration of audio files.*

Data set	**Audio**	**Text**
BUILD, CONSTR	10 hours	Unlimited
BUILD, UNCONSTR	Unlimited	Unlimited
DEV	10 hours	n/a
EVAL	5 hours	n/a

Table 2: *BUILD, DEV, and EVAL data set file and transcribed word counts.*

	BUILD		DEV		EVAL	
Lang.	**Files**	**Words**	**Files**	**Words**	**Files**	**Words**
AMH	122	64,391	123	65,763	64	33,241
CAN	120	96,943	120	95,893	69	50,087
GUA	134	68,984	124	71,285	62	36,199
JAV	122	64,047	122	68,765	62	33,638
KUR	133	82,418	132	77,930	66	38,479
MON	126	90,258	124	90,260	60	44,306
PAS	131	108,509	136	108,713	60	50,693
SOM	132	87,670	126	85,666	66	44,951
TAM	125	70,980	125	71,107	64	36,057
VIE	126	111,952	132	112,029	68	56,048

2.4. Metrics

The primary metric computed on the submitted output was Word Error Rate (WER), as implemented in the sclite tool of the Speech Recognition Scoring Toolkit SCTK available from NIST.[13] WER is computed as the sum of deletion, insertion, and substitution errors in the ASR output compared to a human reference transcription, divided by the total number of words in the human reference transcription:

$$WER = \frac{\#Deletions + \#Insertions + \#Substitutions}{\#ReferenceWords} \tag{1}$$

Character Error Rate (CER) was also computed. CER is calculated in the same way as WER, but at the character level instead of word level.

In addition, participants were required to self-report time and memory resources used by their ASR system(s). The time information was used to compute a run time factor (compared to the real time of the audio data processed) as a secondary metric, while the memory information was to provide the community with information about the resources required to use the ASR system(s).

3. Participation

Interest in the challenge was high. Originally, 28 teams from twelve countries registered to participate. Interest was fairly evenly distributed across the offered languages. In the end, a total of nine teams from five countries participated *fully*, meaning they made at least one valid CONSTR training submission on at least one language's EVAL data set. The participating organizations, their countries, and their team names as used in the results are listed in Table 3.

Table 3: *Participants. Asterisk (*) indicates team did not submit required system description.*

Organization	Country	Team Name
Catskills Research Co.	USA	Catskills
Centre de Recherche Informatique de Montréal	Canada	CRIM
National Sun Yat-sen University	Taiwan	NSYSU-MITLab
Shanghai Jiao Tong University	China	Speechlab-SJTU
*Tal	China	*upteam
Tallinn University of Technology	Estonia	TalTech
Tencent	China	MMT
Tencent, Tsinghua University collaboration	China	TNT
Tsinghua University	China	THUEE

The total number of valid submissions across all participants, languages, and training conditions was 128. The number of teams per language and valid submissions by language and training condition are listed in Table 4.

Table 4: *Participation by language.*

Language	Teams	CONSTR	UNCONSTR
AMH	4	9	1
CAN	6	15	7
GUA	4	9	1
JAV	4	9	1
KUR	5	14	2
MON	6	14	6
PAS	3	7	0
SOM	5	13	1
TAM	4	8	1
VIE	4	10	0

All languages received interest and submissions. Cantonese and Mongolian had the most teams participating (6) and also received the highest number of submissions overall (22 for CAN, 20 for MON). Regarding training conditions, there were more submissions for the required CONSTR training condition than the optional UNCONSTR training condition, for all languages.

4. Results and Analysis

An overview of the OpenASR20 Challenge results is available online.[14] The following synopsis focuses on results for WER and, to allow for more meaningful comparisons, on the CONSTR training condition.

Table 5 lists the best WER score for each team that participated in that language, as well as the CER score for the same submission. Scores are ordered from best (lowest) to worst (highest) WER, by training condition, though late submissions and submissions by teams who did not submit the required system description are flagged as indicated and listed at the bottom of the respective category. Submissions for UNCONSTR training were overall rare and did not exist for all languages. For those languages with UNCONSTR submissions, the best UNCONSTR score was better than the best CONSTR score only in the case of Cantonese and Mongolian.

Table 5: *WER and CER for each team's best WER submission by language and training condition. Asterisk (*) indicates missing system description, dagger (†) indicates late submission; both cases are listed at the bottom of their category.*

Lang.	Condition	Team	WER	CER
AMH	CONSTR	TalTech	0.45	0.34
	CONSTR	THUEE	0.46	0.35
	CONSTR	Speechlab-SJU	1.02	0.89
	CONSTR	*upteam	*1.38	*1.36
AMH	UNCONSTR	Speechlab-SJTU	1.02	0.89
CAN	CONSTR	TNT	0.40	0.35
	CONSTR	THUEE	0.44	0.38
	CONSTR	TalTech	0.45	0.40
	CONSTR	NSYSU-MITLab	0.61	0.56
	CONSTR	Speechlab-SJTU	0.76	0.70
	CONSTR	*upteam	*1.31	*1.33
CAN	UNCONSTR	TNT	0.32	0.26
	UNCONSTR	Speechlab-SJTU	0.76	0.70
GUA	CONSTR	THUEE	0.46	0.42
	CONSTR	TalTech	0.47	0.43
	CONSTR	Speechlab-SJTU	0.99	0.96
	CONSTR	*upteam	*1.21	*1.21
GUA	UNCONSTR	Speechlab-SJTU	0.99	0.96
JAV	CONSTR	THUEE	0.52	0.52
	CONSTR	TalTech	0.54	0.54
	CONSTR	Speechlab-SJTU	0.94	0.94
	CONSTR	*upteam	*1.35	*1.35
JAV	UNCONSTR	Speechlab-SJTU	0.94	0.94
KUR	CONSTR	TalTech	0.65	0.61
	CONSTR	THUEE	0.67	0.62
	CONSTR	CRIM	0.75	0.71
	CONSTR	*upteam	*1.09	*1.08
	CONSTR	Speechlab-SJTU	1.12	1.05
	CONSTR	*upteam	*1.09	*1.08
KUR	UNCONSTR	Speechlab-SJTU	1.12	1.05
MON	CONSTR	THUEE	0.45	0.33
	CONSTR	MMT	0.45	0.33
	CONSTR	TalTech	0.47	0.35
	CONSTR	Speechlab-SJTU	0.97	0.80
	CONSTR	†TNT	†0.45	†0.35
	CONSTR	*upteam	*1.03	*1.00
MON	UNCONSTR	MMT	0.41	0.30
	UNCONSTR	TNT	0.46	0.34
	UNCONSTR	Speechlab-SJTU	0.97	0.80
PAS	CONSTR	TalTech	0.46	0.32
	CONSTR	THUEE	0.49	0.34
	CONSTR	*upteam	*1.37	*1.35
SOM	CONSTR	TalTech	0.59	0.59
	CONSTR	THUEE	0.60	0.60
	CONSTR	Speechlab-SJTU	1.04	1.04
	CONSTR	Catskills	1.14	1.14
	CONSTR	*upteam	*1.23	*.23
SOM	UNCONSTR	Speechlab-SJTU	1.04	1.05
TAM	CONSTR	TalTech	0.65	0.42
	CONSTR	THUEE	0.66	0.44
	CONSTR	Speechlab-SJTU	1.06	0.80
	CONSTR	*upteam	*1.35	*1.32
TAM	UNCONSTR	Speechlab-SJTU	1.06	0.80
VIE	CONSTR	TalTech	0.45	0.41
	CONSTR	THUEE	0.46	0.41
	CONSTR	NSYSU-MITLab	0.75	0.70
	CONSTR	*upteam	*1.41	*1.41

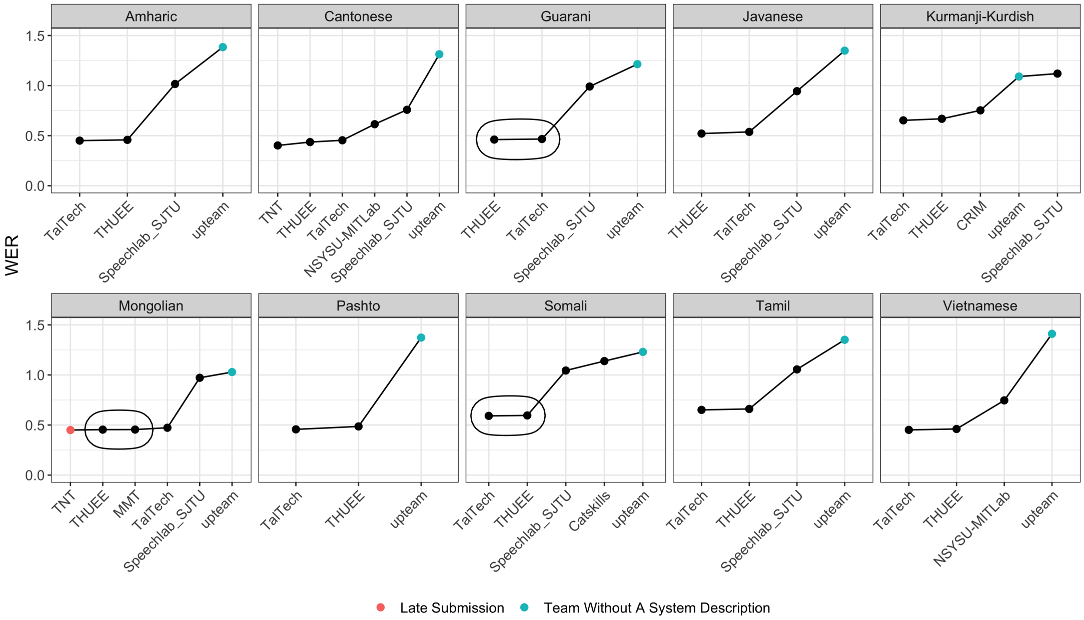


Figure 1: *Best CONSTR WER scores per language and team. Oval enclosing the data points indicates no significant difference at the 95% level.*

The best CONSTR WER scores per team and language were tested for significant differences using Student's t-test. Adjacent data points surrounded by an oval are scores that did not differ significantly according to this test ($p > 0.05$) as shown in Figure 1.

WER was overall high. The best (lowest) results per language range from 0.40 (Cantonese) to 0.65 (Kurmanji Kurdish), with many of the languages in between close to the 0.5 mark for the best result. As mentioned in the Data section, some metadata was available, but some of these factors only had a small amount of data - not enough for comparison. We focused on those with adequate samples (at least 20 audio files). Using the best CONSTR submission for each language, we explored the effect of dialect and gender on WER. For the dialect distinction, not every case had enough data in each category for a meaningful comparison. For those that did, no significant differences between WER for different dialects were found. For the gender distinction, a significant difference ($p < 0.05$) between WER for male vs. female speech was found only for Javanese and Pashto; in both cases the female speech resulted in significantly better WER than the male speech.

5. Discussion and Conclusions

The results indicate that ASR for low-resource languages, and in particular paired with a challenging data type such as conversational telephone speech, remains a difficult problem, as evidenced by the worse WER scores compared to more widely studied languages with large amounts of training data available. Some of the languages tested in OpenASR20 (CAN, KUR, PAS, TAM, VIE) were tested in the aforementioned Babel program from 2013-2016 as well, sourcing materials from the same larger data sets. This allows for some limited comparability, although the evaluation data sets used were not identical, and notably the training data was not limited as severely in Babel as in OpenASR20. While the results were not directly comparable, the best scores achieved in OpenASR20 were close to those achieved in Babel years before with more training data, and in some cases better. This indicates potential progress in that similar results can now be achieved with less resources.

We plan on performing error analyses in relation to challenges specific to the different language data, as well as, in collaboration with participants, in relation to training data usage.

While 28 teams originally registered for the challenge, only nine made valid submissions on at least one language's EVAL set. It will be useful to determine the reasons for this drop, and how to lower the barriers to entry to encourage wider participation. One consideration to become more accessible is to provide a computing platform with the infrastructure for participants to run their systems, instead of them having to rely on their own, which may vary widely between organizations. In addition, it may be useful to examine the usefulness of the UNCONSTR training condition with its low participation and minimal or no performance gain; its presence may represent another factor discouraging wider participation.

6. 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. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

7. References

[1] D. S. Pallett, "The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program," *Speech Communication*, vol. 37, no. 1, pp. 3–14, May 2002.

[2] A. F. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *2007 IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, 2007, pp. 1–7.

[3] NIST. (2009) Rich transcription evaluation. [Online]. Available: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, Jan. 2014.

[5] D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds., *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. Marseille, France: European Language Resources association, May 2020.

[6] IARPA. (2016) Babel. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/babel

[7] NIST. (2021) OpenASR Challenge. [Online]. Available: https://www.nist.gov/itl/iad/mig/openasr-challenge

[8] IARPA. (2021) MATERIAL. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/material

[9] NIST. (2020) OpenSAT. [Online]. Available: https://www.nist.gov/itl/iad/mig/opensat

[10] ——. (2013) IARPA Babel data specifications for performers. [Online]. Available: https://www.nist.gov/system/files/documents/itl/iad/mig/IARPA_Babel_Performer-Specification-08262013.pdf

[11] I. Zavorin, A. Bills, C. Corey, M. Morrison, A. Tong, and R. Tong, "Corpora for cross-language information retrieval in six less-resourced languages," in *Proceedings of the Workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. Marseille, France: European Language Resources Association, May 2020, pp. 7–13.

[12] NIST. (2020) OpenASR20 Challenge evaluation plan. [Online]. Available: https://www.nist.gov/document/openasr20-challenge-evaluation-plan

[13] ——. (2018) SCTK, the NIST scoring toolkit. [Online]. Available: https://github.com/usnistgov/sctk

[14] ——. (2021) OpenASR20 Challenge results. [Online]. Available: https://www.nist.gov/itl/iad/mig/openasr20-challenge-results

Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data across NASA Apollo Audio

Aditya Joglekar[1], *Seyed Omid Sadjadi*[2], *Meena Chandra-Shekar*[1],
Christopher Cieri[3], *John H.L. Hansen*[1]

[1]Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
The University of Texas at Dallas (UTD), Richardson, Texas, USA
[2]NIST ITL/IAD/MIG, Gaithersburg MD, USA
[3]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
{aditya.joglekar, meena.chandrashekar, john.hansen}@utdallas.edu[1],
omid.sadjadi@nist.gov[2], ccieri@ldc.upenn.edu[3]

Abstract

The Fearless Steps Challenge (FSC) initiative was designed to host a series of progressively complex tasks to promote advanced speech research across naturalistic "Big Data" corpora. The Center for Robuts Speech Systems at UT-Dallas in collaboration with the National Institute of Standards and Technology (NIST) and Linguistic Data Consortium (LDC) conducted Phase-3 of the FSC series (FSC P3), with a focus on motivating speech and language technology (SLT) system generalizability across channel and mission diversity under the same training conditions as in Phase-2. The FSC P3 introduced 10 hours of a previously unseen channel audio from Apollo-11 and 5 hours of novel audio from Apollo-13 to be evaluated over previously established and newly introduced SLT tasks with streamlined tracks. This paper presents an overview of the newly introduced conversational analysis tracks, Apollo-13 data, and analysis of system performance for matched and mismatched challenge conditions. We also discuss the Phase-3 challenge results and evolution of system performance across the three Phases, and next steps in the Challenge Series.

Index Terms: NASA Apollo, speech activity detection, speaker diarization, speaker identification, speech recognition, conversational analysis.

1. Introduction

The importance of naturalistic datasets in the advent of a technology revolution led by deep neural networks has become paramount. With deep learning system performance linearly scaling with the amount of data provided, naturalistic "Big Data" corpora have become a benchmark in determining a competitive edge in the domain of artificial intelligence (AI). Developing good quality naturalistic datasets is challenging [1, 2, 3, 4].

There are additional challenges in developing data with rich information content in multiple SLT domains [5, 6]. The NASA Apollo missions data is a collection of 150,000+ hours of audio with unprompted multi-party conversations recorded over 30 channels. The preserved data contains over 450 personnel in constant air-to-air, air-to-ground, and ground-to-ground communication working collaboratively to solve time-critical challenges. Speech and natural language systems can significantly benefit from this preserved data and vice versa. The Fearless Steps Challenge (FSC) series has led the efforts in promoting such system development by annotating a small portion of this Apollo corpus (115 hours) and establishing challenge tasks to benchmark SLT systems [6, 7, 8].

The goal of FSC Phase-3 (FSC P3) is to assess the ability of speech and language systems in providing consistent performance across channel, noise, speech, speaker, and semantic variabilities. The FSC P3 evaluation set provides such a test platform, sourcing its data from multiple channels from Apollo-11 and Apollo-13 missions [9, 10, 11].

FSC P3 was conducted in collaboration with the National Institute of Standards and Technology (NIST) and Linguistic Data Consortium (LDC) from February through March of 2021. Nine participating organizations, with 14 task-specific teams contributed 235 system submissions. Phase-3 evaluation reported state-of-the-art (SOTA) results on 3 of the 8 challenge tracks. The current edition of the FSC was run entirely online using the NIST OpenSAT evaluation platform (`https://sat.nist.gov/fsc3`). The web platform supported a variety of services including evaluation registration, data distribution, system output submission, submission validation, scoring, and system description/presentation uploads [5, 12, 13].

2. Challenge Tasks

The NASA Apollo Missions audio data were recorded on 30-channel analog tapes ranging from 14 to 17 hours in duration. Large data streams of this nature are often a hindrance in effective development of SLT systems. In an effort to streamline such development, 30-minute audio chunks were annotated from the most high-impact events in Apollo 11 and Apollo 13 missions. These chunks have been presented in the FSC corpus as "audio streams". These streams of undiarized audio are supplemented by annotation files with diarized labels. Essentially, audio streams for Training (*Train*) and Development (*Dev*) sets provided to challenge participants contain a single markup-styled annotation file per stream. Each annotation file contains a ground-truth label or transcription per single-speaker utterance. Participants are expected to diarize the Evaluation (*Eval*) set audio streams in addition to the primary task. This requires using multiple systems to process downstream tasks, often propagating error at each functional block. The FSC P2 established the necessity for developing separate speaker diarization (SD) and automatic speech recognition (ASR) tracks to streamline the development of effective clustering and acoustic models/language models respectively [3, 4, 14].

This format has been extended for the FSC P3, which provides separate tracks for diarized "audio segments" in the speaker diarization (SD), automatic speech recognition (ASR), and conversational analysis (*CONV*) tasks. The entire list of tasks and

Joglekar, Aditya; Sadjadi, Omid; Chandra-Shekar, Meena; Cieri, Christopher; Hansen, John. "Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data across NASA Apollo Audio." Presented at INTERSPEECH 2021: Special Session on OpenASR20 (Open Automatic Speech Recognition) and Low Resource ASR Development, Brno, CZ. August 30, 2021 - September 03, 2021.

tracks available for the FSC P3 are presented below. The five original channels (*FD*, *MOCR*, *EECOM*, *GNC*, *NTWK*) used in the FSC P1 and P2 have been preserved in Train and Dev sets, with no additions to the data-sets [6, 7, 8].

- *TASK 1*: Speech Activity Detection ***(SAD)***
- *TASK 2*: Speaker Identification ***(SID)***
- *TASK 3*: Speaker Diarization
  - ○ (3.a.) *Track 1*: using system SAD ***(SD_track1)***
  - ○ (3.b.) *Track 2*: using reference SAD ***(SD_track2)***
- *TASK 4*: Automatic Speech Recognition
  - ○ (4.a.) Track 1: using system SAD ***(ASR_track1)***
  - ○ (4.b.) Track 2: using diarized audio ***(ASR_track2)***
- *TASK 5*: **Conversational Analysis**
  - ○ (5.a.) Track 1: using system SAD ***(CONV_track1)***
  - ○ (5.b.) Track 2: using diarized audio ***(CONV_track2)***

Tasks established in previous challenges are still core speech tasks, and do not directly benefit spoken language understanding (SLU) of the multi-party conversations. With SOTA word error rates (WER) as high as 24%, SLU systems cannot be expected to effectively extract meaningful information regarding topic, sentiment, emotion, or semantic context. This effect was observed in the FSC P1 sentiment detection task. Accordingly, a new task with separate tracks was created with an aim to identify key conversational moments in the data [9, 15, 16, 17].

2.1. Conversational Analysis

Methodologies to identify salient events in continuous audio streams can significantly reduce the cost of information retrieval [18, 19, 20, 21]. The significance of such methodologies is more pronounced for the Apollo Missions, which can have intermittent time-critical events, followed by large periods of inactivity or normal conversations. Identification of such "Hotspots" in over 150,000 hours of audio data can help both STEM and non-STEM researchers in efficient retrieval and analysis of high value content. These hotspot events are essentially an extractive summarization of conversations between Mission Control personnel. A total of 25 conversational cues critical to successful deployment of the Apollo Missions were identified as conversational "Hotspots" and presented as diarized segments for a classification task termed "Hotspot Detection" (CONV_track2). The task of finding salient events and providing accurate extractive summarization in continuous audio streams is presented as a separate track for the FSC P3 (CONV_track1).

2.2. Challenge Deployment

The NIST OpenSAT web platform (`https://sat.nist.gov`) was used to conduct the FSC P3. Participants were allowed to download Train, Dev, and Eval sets after agreeing to the terms and conditions of the Challenge. The scoring toolkit developed for the FSC P2 was used in the validation and scoring back-end for the platform. Participants were provided with basic analytics for their submissions to assist in their system development efforts. Every team was allocated a single submission slot per task/track with multiple submissions (up to 3 per day) allowed to update the system performance.

2.3. Performance Metrics

The performance metrics and conditions for the FSC P3 were largely the same as the conditions for the FSC P2. A NIST defined detection cost function (DCF) measure was used for scoring the speech activity detection (SAD) task, with a forgiveness collar of *0.25* seconds, which was reduced from the collar duration of *0.5* seconds for the FSC P2. Both tracks for SD and ASR were evaluated using diarization error rate (DER) and word error rate (WER) for the same testing conditions as the FSC P2. Speaker Identification (SID) task performance metric was updated from Top-5 % Accuracy (Top-5 Acc.) to Top-3 % Accuracy (Top-3 Acc.). The newly introduced task tracks CONV track-1 and track-2 were evaluated using separate metrics. Track-2 using diarized segments was evaluated using Top-3 % Accuracy, and track-1 using audio streams was tested using the measure Jaccard error rate (JER) [22, 23, 24]. The Jaccard index has been traditionally used in image segmentation and more recently in speaker diarization in the DIHARD Challenge series [22, 25]. JER as an initial measure provides a good representation of diarizing and identifying conversational labels for the diarized segments. For each reference speaker *ref* the speaker-specific JER$_{ref}$ is computed as:

$$\text{JER}_{\text{ref}}(\%) = \left(\frac{FA + Miss}{Total}\right) \times 100, \qquad (1)$$

where

- *Total* is the total reference speaker time; that is, the sum of the durations of all reference speaker,
- *FA* is the total system speaker time not attributed to a reference speaker,
- *Miss* is the total reference speaker time not attributed to a system speaker.

The JER metric for CONV track1 could be replaced with a different metric for future Phases depending on the evolution of the data and labels developed for this task.

3. Data

The amount of raw unlabelled data generated in the world has been exponentially increasing compared to the available ground-truth annotated data. A key feature of robust SLT systems is their ability to adapt to such real-world data with varying acoustic characteristics, in low-resource training conditions.

3.1. Unseen Channel & Mission

The operations and propulsion (OPS&PRO) channel from Apollo-11 was annotated for testing system performance over unseen noise characteristics. It provides a good basis for testing multiple tasks, since the content of speech due to the technical nature of the channel will be different from the other Apollo-11 channels which have more general mission status related conversations. A total of 5 hours of Apollo-13 data was selected from three separate channels, providing additional variability in channel noise, air-to-ground communication noise, different speakers, and speech conversations of a very different nature compared to the conversations seen in Apollo-11.

3.2. General Statistics

Table 3 summarizes the overall statistics for the audio streams data. The key feature we notice is the variability in the Eval set even for basic analysis parameters. Tables 1 and 2 present

Table 1: *General statistics for the SID task. The mean, median, minimum, and maximum values for cumulative speaker durations, and individual speaker utterances are all expressed in seconds [7]*

Data set	**# Spkrs**	**Spkr. Duration (s)**			**Spkr. Utterances (s)**		
		mean	**median**	**(min , max)**	**mean**	**(min , max)**	**total**
Train	218	505.5	106.7	(6.89 , 11254.36)	4.03	(1.84 , 16.95)	27336
Dev	218	118.1	24.2	(3.13 , 2596.18)	4.04	(1.78 , 16.95)	6373
Eval	218	264.3	38.2	(3.19 , 5834.46)	4.09	(1.8 , 16.22)	14077

Table 2: *General statistics for the CONV_track2 task. The mean, median, minimum, and maximum values for cumulative label durations, and individual labels are all expressed in seconds.*

Data set	**# Hotspots**	**Lab. Duration (s)**			**Lab. Utterances (s)**		
		mean	**median**	**(min , max)**	**mean**	**(min , max)**	**total**
Train	25	1546.8	796.4	(193.26 , 6274.85)	2.4	(0.5 , 30.4)	16059
Dev	25	464.8	233.7	(45.11 , 2626.05)	2.5	(0.5 , 33.2)	4662
Eval	25	976.3	435.9	(59.05 , 8036.36)	2.6	(0.5 , 29.96)	9360

Table 3: *Overall Statistics of audio streams for the FSC P3. The mean, min, and max values are expressed in seconds.*

	Train	**Dev**	**Eval**			
			A-11	A-13	Unseen	**Total**
# Streams	125	30	40	10	18	68
Dur. (hrs)	63.5	15.3	20.1	5	9.1	34.4
% Speech	29.4	32.5	34.4	33.4	37.3	35
#Spkrs/Stream	19	24	20	13	25	20

Table 4: *Duration Statistics of audio segments for ASR_track2. The mean, min, and max values are expressed in seconds [7]*

Data set	**Segments**	**Utterance Duration (s)**		
		mean	**min**	**max**
Train	35,473	2.85	0.10	70.37
Dev	9,203	2.97	0.12	67.39
Eval	21,846	2.98	0.10	162.75

statistics on the diarized segments for the SID and CONV_track2 tasks, while Table 4 provides insight into the distribution of segments in the Train, Dev and Eval sets.

3.3. Evaluation Set Variability

The unseen channel and mission data present in the evaluation set are set up in a blind format. Participants were not provided with mission or channel labels during the evaluation phase, thereby making it possible to evaluate systems for their generalizability to unseen data conditions.

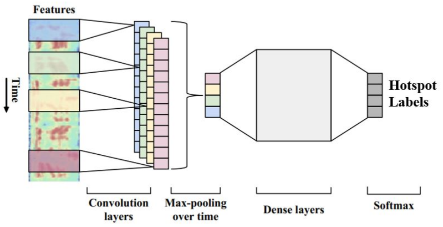


Figure 1: *CNN-Saliency network illustration, used as baseline system for both tracks of Conversational Analysis [26]*

4. Results

In this section, we analyze baseline performance for all tasks, and evaluate the system generalizability factor for submissions from all challenge participants.

4.1. Baseline Results

All the original FSC2 baseline systems were used without modifications for evaluating the FSC3 Eval set [27, 28, 29] with the exception of the extractive summarization tasks (CONV track1 and track2). The baseline results for all tasks are reported in Table 5. We notice a degradation in performance on the FSC3 Eval sets for all tasks/tracks, caused by the addition of the unseen channel and mission data. Since most of the baseline systems are unsupervised and rely on core acoustic features [30, 31], their degraded performance indicates added acoustic complexity in the Eval set due to the unseen mission and channel inclusion.

The baseline system used for both tracks of the conversational analysis task was originally developed for emotion detection[26]. The convolutional network illustrated in Figure 1 is designed to aggregate high level context over time (in this case, log-Mel-Filterbank feature frames) to generate a single summarization, which fits well with the structure of this task.

4.2. Best Systems Comparison

Table 6 provides a comparative illustration of the improvements in SOTA for Phase-2 and Phase-3[32, 33, 34, 35]. We report significant improvements in FSC3 Top system performance over FSC2 for SID task, ASR_track2 and both SD tracks. However, the SOTA for SAD and ASR_track1 tasks are maintained from FSC2. Comparisons to the FSC2 systems have been conducted

Table 5: *Baseline Results for FSC P3 Development and Evaluation Sets [7]*

Fearless Steps Phase-3 Baseline Results			
Task	**Metric**	**Dev (%)**	**Eval (%)**
SAD	DCF	12.84	15.16
SD_track1	DER	79.72	88.27
SD_track2	DER	68.68	77.91
SID	Top-3 Acc.	75.20	72.46
ASR_track1	WER	83.80	92.3
ASR_track2	WER	80.50	86.4
CONV_track1	JER	58.6	71.5
CONV_track2	Top-3 Acc	67.1	54.2

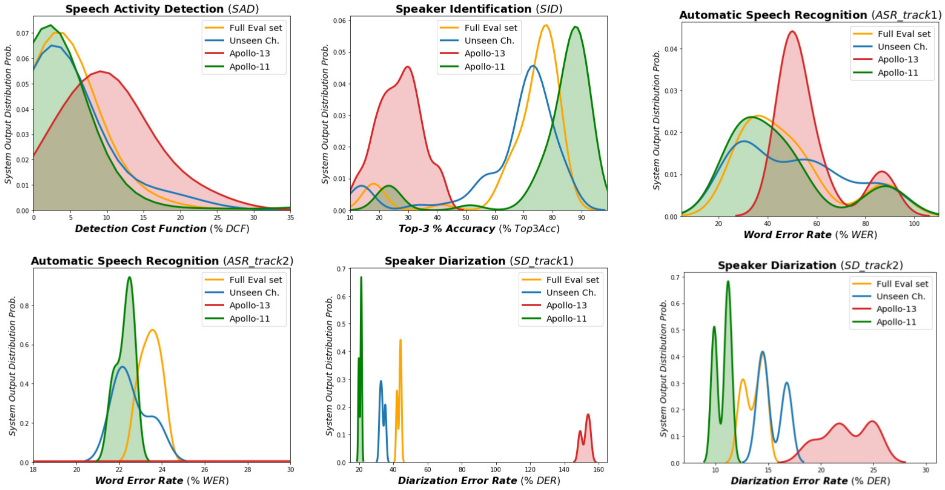


Figure 2: *Distribution of system submission results between the Apollo-11, Unseen Channel, Apollo-13 segments/streams of the Eval set.*

Table 6: *Comparison of the best systems developed for all FSC P2 and P3 challenge tasks.* FS-3 *(in bold) represents system performance over the entire blind set provided for participants. Relative improvement of top-ranked system per task in the FSC P3 seen channel data* (FSC3-A11) *(underlined) over FSC P2 evaluation set is illustrated.* FSC3-A13 *and* FSC3-UnkCh *represent the top system performance for the Unseen Mission and Unseen channel data respectively.*

Comparison of Best System Submissions on FSC3 Eval set (and sub-sets)						
Task	**FSC2 (%)**	**FSC3 (%)**	**FSC3-A11 (%)**	**FSC3-A13 (%)**	**FSC3-UnkCh (%)**	**Rel. Imp. (%)**
SAD	1.07	**1.47**	1.16	2.37	1.67	*-7.57 %*
SID	92.39	**83.27**	93.26	23.77	85.16	**12.9 %**
SD_track1	28.85	**42.20**	19.92	149.1	32.33	**30.95 %**
SD_track2	26.55	**12.32**	9.82	19.05	14.15	**53.59 %**
ASR_track1	24.01	**29.96**	26.87	47.94	26.82	*-11.9 %*
ASR_track2	24.26	**22.85**	21.69	100	21.82	**10.59 %**

over the same data evaluated by participants in FSC2. We also observe that the top systems for every task/track were able to adapt efficiently to the unseen channel data.

4.3. System Generalizability

Figure 2 displays a comparative performance of all system submissions across the Apollo-11, Unseen Channel, and Apollo-13 sub-sets of the Eval set. We observe that while the top systems were able to generalize well to the unseen mission and channel data, a majority of the systems had degraded performance. The unseen channel performance for all tasks was seen to be closely related to the seen channel data, and in some cases even showed improved performance. Based on this analysis, we hypothesize that system generalizability is a key component in developing systems for the remaining Apollo Missions.

5. Discussion

We notice that system generalizability was positively correlated with a better overall performance. We report that the systems competing in the FSC P3 showed improved performance for channel variabilities, but significantly lacked the ability to generalize to the unseen mission data. It should be noted that the imbalance in the Eval sub-sets could have caused some bias in the overall submission results. We plan to create and provide equally weighted channel and mission data for the next challenge.

6. Conclusions

Through the Phase-3 of the FSC, we introduced a new challenge task that aims to extract high level context from conversations. We tested system capability to generalize for previously unseen channel and mission variability. In the next Phase of the Challenge we plan to extend the Training and Development datasets to include more data from the Apollo missions 8 and 10. In conclusion, we assert the need for further development in SLT systems for naturalistic data. Future efforts in the Fearless Steps Challenge series will increasingly involve the goal of developing systems that are highly adaptable and robust to out-of-domain data.

7. Acknowledgements

This project was supported in part by AFRL under contract FA8750-15-1-0205, NSF-CISE Project 1219130, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H. L. Hansen. We would also like to thank Tatiana Korelsky and the National Science Foundation (NSF) for their support on this scientific and historical project. A special thanks to Katelyn Foxworth (CRSS Transcription Team) for leading the ground-truth development efforts on the FSC P3 Challenge Corpus.

8. References

[1] 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, p. 101032, 2020.

[2] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[3] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE ASRU Workshop*, 2015, pp. 547–554.

[4] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*, 2018, pp. 1561–1565.

[5] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 224–230.

[6] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. INTERSPEECH*, 2018, pp. 2758–2762.

[7] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "Fearless steps challenge (FS-2): Supervised learning with massive naturalistic Apollo data," in *Proc. INTERSPEECH*, 2020, pp. 2617–2621.

[8] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," in *Proc. INTERSPEECH*, 2019, pp. 1851–1855.

[9] A. Joglekar and J. H. Hansen, "Fearless steps, NASA's first heroes: Conversational speech analysis of the apollo-11 mission control personnel," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2956–2956, 2019.

[10] J. C. Gorman, P. W. Foltz, P. A. Kiekel, M. J. Martin, and N. J. Cooke, "Evaluation of latent semantic analysis-based measures of team communications content," in *Proceedings of the Human Factors and Ergonomics Society annual meeting*, vol. 47, no. 3. SAGE Publications Sage CA: Los Angeles, CA, 2003, pp. 424–428.

[11] P. W. Foltz, D. Laham, and M. Derr, "Automated speech recognition for modeling team performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, no. 4. SAGE Publications Sage CA: Los Angeles, CA, 2003, pp. 673–677.

[12] 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*, 2017, pp. 1353–1357.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *STIN*, vol. 93, p. 27403, 1993.

[14] A. Sangwan, L. Kaushik, C. Yu, J. H. Hansen, and D. W. Oard, "'houston, we have a solution' : Using NASA Apollo program to advance speech and language processing technology," in *Proc. INTERSPEECH*, 2013, pp. 1135–1139.

[15] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1688–1700, 2014.

[16] F. Valente and A. Vinciarelli, "Language-independent socio-emotional role recognition in the AMI meetings corpus," in *Proc. INTERSPEECH*, 2011.

[17] S. Raaijmakers, K. P. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 466–474.

[18] S. Somasundaran, J. Ruppenhofer, and J. Wiebe, "Detecting arguing and sentiment in meetings," in *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 26–34.

[19] J. Carletta and J. Kilgour, "The NITE XML toolkit meets the ICSI meeting corpus: Import, annotation, and browsing," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 111–121.

[20] A. Dielmann and S. Renals, "DBN based joint dialogue act recognition of multiparty meetings," in *Proc. IEEE ICASSP*, vol. 4, 2007, pp. IV–133.

[21] C. Lai, J. Carletta, S. Renals, K. Evanini, and K. Zechner, "Detecting summarization hot spots in meetings using group level involvement and turn-taking features." in *Proc. INTERSPEECH*, 2013, pp. 2723–2727.

[22] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," 2018.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU Workshop*, 2011.

[24] "NIST rich transcription spring 2003 evaluation," https://catalog.ldc.upenn.edu/LDC2007S10, accessed: 2019-03-01.

[25] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[26] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. IEEE ICASSP*, 2017, pp. 2741–2745.

[27] Z.-H. Tan, N. Dehak *et al.*, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, 2020.

[28] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[29] H. Dubey, A. Sangwan, and J. H. Hansen, "Robust speaker clustering using mixtures of von Mises-Fisher distributions for naturalistic audio streams," in *Proc. INTERSPEECH*, 2018, pp. 3603–3607.

[30] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, March 2013.

[31] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[32] B. Sharma, R. K. Das, and H. Li, "Multi-level adaptive speech activity detector for speech in naturalistic environments," *Proc. INTERSPEECH*, pp. 2015–2019, 2019.

[33] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Two-dimensional convolutional recurrent neural networks for speech activity detection," in *Proc. INTERSPEECH*, 2019.

[34] A. Gorin, D. Kulko, S. Grima, and A. Glasman, ""This is Houston. Say again, please". the Behavox system for the Apollo-11 fearless steps challenge (Phase II)," *arXiv preprint arXiv:2008.01504*, 2020.

[35] Q. Lin and M. L. Tingle Li, "The DKU speech activity detection and speaker identification systems for fearless steps challenge Phase-02," *Proc. INTERSPEECH*, pp. 2607–2611, 2020.

A System for Validating Resistive Neural Network Prototypes

Brian Hoskins, Mitchell Fream, Matthew Daniels, Jonathan Goodwill, Advait Madhavan, Jabez McClelland
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

Osama Yousuf, Gina Adam
George Washington University
Washington, DC, USA

Wen Ma, Tung Hoang, Mark Branstad, Muqing Liu, Rasmus Madsen, Martin Lueker-Boden
Western Digital Research
San Jose, California, USA

ABSTRACT

Building prototypes of heterogeneous hardware systems based on emerging electronic, magnetic, and photonic devices is an important area of research. The novel implementation of these systems for artificial intelligence poses new and unforeseen challenges in mixed signal data acquisition, hyperparameter optimization, and hardware co-processing. Many emerging devices exhibit unpredictable and stochastic behavior as well as poorly repeatable hysteretic effects or performance degradation. Dealing with these device challenges on top of more traditional hardware problems, like quantization errors, timing constraints, and even hardware and software bugs is an enterprise fraught with pitfalls. Equally important to the construction of the physical prototype is the co-development and integration of a design verification framework that can extensibly allow for predictable behavior of not only the entire system but also all of its parts in a modular way, allowing for seamless integration in both simulation and implementation. This work discusses Daffodil-lib, a Python based prototyping framework which, from hardware to software, enables everything from a script-based simulation to a compiled hardware-timed experiment, to everything in between with no syntactical changes for the end user.

CCS CONCEPTS

• **Hardware → Emerging tools and methodologies**; **Memory and dense storage**; **Emerging simulation**.

KEYWORDS

design verification, neural networks, prototyping, hardware

ACM Reference Format:
Brian Hoskins, Mitchell Fream, Matthew Daniels, Jonathan Goodwill, Advait Madhavan, Jabez McClelland, Osama Yousuf, Gina Adam, and Wen Ma, Tung Hoang, Mark Branstad, Muqing Liu, Rasmus Madsen, Martin Lueker-Boden. 2021. A System for Validating Resistive Neural Network Prototypes. In *International Conference on Neuromorphic Systems 2021 (ICONS 2021), July 27–29, 2021, Knoxville, TN, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477145.3477260

1 INTRODUCTION AND BACKGROUND

Increasing numbers of novel neuromorphic prototypes are becoming available. These include traditional silicon technologies that span from implementing novel architectures, such as Loihi, to more conventional deep neural networks [9, 18]. These conventional hardware systems benefit from a long history of design verification, which includes industry practice in classic digital system design, as well as practice in software development [4, 11, 12, 20]. Traditionally, design verification took place within the space of commercial electronic design automation tools, but now a growing number of open source tools – especially those based on Verilator[1], a tool for compiling ("verilating") Verilog into accessible C++ libraries – have facilitated an explosion of alternatives. These alternatives have primarily been used for hardware description and verification using high-level languages [3, 8, 13, 17, 22]. These recent advancements have expanded the utility and interoperability of digital system design with high-level modeling, making it easier to simulate the interaction of a digital system with the world or an analog system.

In parallel to these advancements, interest has grown in the development of analog systems based on both conventional silicon as well as on unconventional devices like resistive memories (ReRAM), phase change memories (PCM), magnetic tunnel junctions (MTJs), and even photonic modulators [1, 5, 10, 14]. In addition to an expanding number of prototypes, there have also been new high-level modeling and simulation tools being developed, but these have only just begun to grow into design verification frameworks [2, 6, 19, 21]. In general, design verification for analog hardware systems is not as mature as it is for digital systems [7], and this is doubly the case for systems that must also incorporate complex dynamics such as weight updates or spike timing dependent plasticity [5].

Consequently, as prototypes develop, it is important to evolve in parallel the tools, methodologies, and software frameworks necessary to ensure that the system functions correctly, accurately predicts a system's performance, and encompasses the total span of hardware/model isomorphisms required for a prototype to be directly translated into an integrated system on a chip. To that end, we introduce Daffodil, a modular end-to-end system capable of simulating and executing experiments on arrays of up to 20,000 resistive devices. The system is composed of an integrated circuit, a mixed-signal daughterboard, a field-programmable gate array (FPGA) development board, and a software framework including a compiled CPU, embedded Linux distribution, FPGA hardware drivers, and Python-based application programming interfaces (APIs).

[1] https://www.veripool.org/verilator/

Hoskins, Brian; Fream, Mitchell; Daniels, Matthew; Goodwill, Jonathan; Madhavan, Advait; McClelland, Jabez J.; Yousuf, Osama; Adam, Gina; Ma, Wen; Liu, Muqing ; Madsen, Rasmus; Lueker-Boden, Martin. "A System for Validating Resistive Neural Network Prototypes." Presented at International Conference on Neuromorphic Systems 2021, Knoxville, TN, US. July 27, 2021 - July 29, 2021.

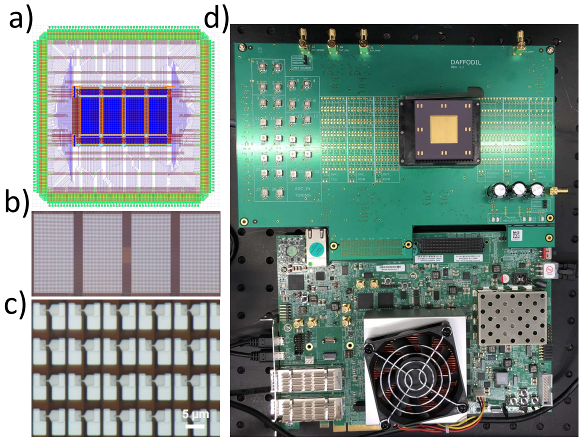


Figure 1: Details of prototyping platform. a) GDS file of the 20,000 device chip. b) 20,000 devices fabricated on the chip. c) Micrograph of a smaller subarray. d) Custom ADC/DAC daughterboard and packaged chip interfacing to an FPGA unit.

The system design leans heavily on conforming to modular interfaces, software-hardware polymorphisms within the APIs, and functionally exact reconstructions of the hardware operations in simulation [12]. The goal of the verification framework is not only to model the system's behavior, but also to model the behavior of programming and algorithmic mistakes so that the simulation predicts even unintended side effects.

2 HARDWARE AND OPERATING SYSTEM

2.1 Integrated Circuit

The integrated circuit, called the Daffodil Chip, is a flexible-interface platform specifically designed for the research and development of two-terminal resistive memory and selector devices. It is designed in a commercial 3.3 V, 180 nm complementary metal oxide semiconductor (CMOS) process and contains via and access points for direct integration of 20,000 resistive devices in a research foundry. The chip is accessed by 403 pads, including three digital CMOS logic configuration pads, 50 gate-access pads, 50 column-access pads, and 200 row-access pads for each side of a double connected row. The remaining pads are for electrical power. Two of the digital pads are used for accessing internal 4:1 multiplexers, which allow for access to one of four internal 50x100 2T-1R arrays. Consequently, the maximum number of devices accessed at a time is 5,000.

2.2 Mixed Signal Daughterboard

The chip interfaces with a mixed-signal daughterboard, which we call the Daffodil Board. The Daffodil Board is designed to access any of the up to 32 unique subarrays, or *kernels*, within the 20,000 array chip. Each kernel, which can be considered as a 25 × 25 2T-1R crossbar, is accessed with 75 unique digital-to-analog converter (DAC) channels, giving each of the rows, columns, and gate columns a unique bias. For read operations, the row or column DACs can be replaced with tunable transimpedance amplifiers feeding to an array of 25 analog to digital converter (ADC) channels. Negative current biases can be emulated in the transimpedance amplifiers by raising their reference inputs from ground using one of the DAC channels. For diagnostic purposes, the rows, columns, and gates can also be globally or individually connected to ground or an external source through coaxial connectors. Logically, the board's selection amongst the 32 kernels is mediated by five external CMOS logic signals. Individual rows and columns can be activated by any of 50 unique CMOS logic signals. Three sets of three CMOS logic signals configure the DAC, ADC, ground, external configurations, or disable the access configuration multiplexers completely. All the 96 CMOS logic configuration signals, the ADC and DAC serial-peripheral interface (SPI) lines, clock signals, and the board power are sourced from a 400-pin FPGA mezzanine connector (FMC).

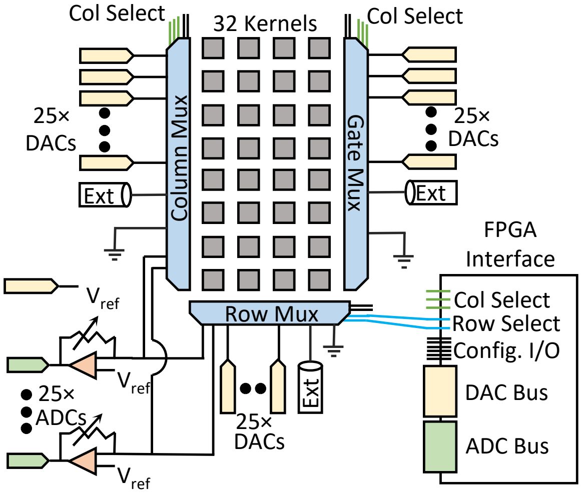


Figure 2: Logical diagram of the prototyping architecture. Five CMOS logic signals access any of 32 ReRAM kernels. Each of the 25 rows, columns, and gates can be independently biased with DACs, accessed externally, or grounded. Transimpedance amplifiers allow for readout of the array.

3 SOFTWARE

3.1 Operating System and Drivers

The Daffodil Board is designed to be plugged into a commercial FPGA development board to act as the online host. The FPGA host operates primarily from a compiled Xilinx Microblaze[2] soft-processor which interfaces to the board drivers and to a superhost Linux PC via Ethernet. The processor operates with a custom embedded Linux distribution, Daffodil Linux, built from the Yocto Project toolset[3]. All of the digital control signals on the Daffodil Board, driven by 1.8V CMOS logic, are accessible as general purpose input/output (GPIO) signals from the Daffodil Linux operating

[2] Certain commercial processes and software are identified in this article to foster understanding, and. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the processes and software identified are necessarily the best available for the purpose.
[3] https://www.yoctoproject.org/

system and can be asserted individually. This allows for the DACs and ADCs to be controlled using off-the-shelf drivers.

In addition to the hardware drivers, the Daffodil Linux distribution contains hooks to call any hardware compiled function on the FPGA. These functions can control any precise, hardware timed operation including pulse generation, read operations, kernel selection, random number generation, random number generation, or algorithmic coprocessing of batch and gradient information.

3.2 Daffodil-lib Base and Simulation Classes

Once the physical structures of the Daffodil Chip and Board are set and all driver definitions are encoded, the Daffodil library (Daffodil-lib) operates as the system's simulation engine, control software, and design verification framework. Written in Python, the library is engineered to describe the board operation in a predictive and a descriptive manner. It is descriptive in the sense that the code written under Daffodil-lib must set the relevant 1.8V logic lines, modeled by integers or Booleans, to select a kernel or specify inputs/outputs—but it is also predictive in the sense that the system must check what the existing logic lines are set to before executing a simulation of the expected behavior. The key action of the predictive behaviors is to set the correct simulated values into the DACs and ADCs as well as the correct voltages and currents in the ReRAM memory.

In the case of the DACs and ADCs, each is modeled with a part class that reproduces the specified transfer functions from the digital to the analog domain, which, in the case of the Daffodil board, is 12-bit digitization. Consequently, any mapping from floating point abstract layers down to the analog layers are automatically rounded and any communication from analog to the abstract neural net layers are also digitized.

The ReRAM memory model interface exposes only input voltages, output currents, and time values as accessible parameters. Consequently, the internals of the model could be anything: a Python model, a C model, or a full SPICE simulation of the system including noise, or even commands to execute operations on a real device. In this section, we discuss both a Python simulation class and a physical operation class. Our implemented ReRAM model and interface does not distinguish between inference and learning operations, and so automatically checks for the possibility of device conductance changes based on the inputted biases. One concrete class descended from the base class is our Python model, named the *simulation class.* As a child of the base class, the simulation class retains the descriptive ADC and DAC models, e.g., it considers the 12-bit. Additionally, the class implements its predictive behavior through a Python wrapped ReRAM model.

The key operation of the simulation is the *event* operation, which maps the correct biases to and from the DACs and ADCs to the ReRAM memory model. Using the input pulse time duration, the simulator predicts the ReRAM behavior, including operationally erroneous behaviors that a user might invoke in higher level code. For example, an inference operation with too-high bias is indistinguishable from a write operation, and improperly set gate biases will not generate any current. Improperly set CMOS logic lines would generate the wrong behavior in the simulation exactly as they would in the experiment. The potential to capture mistakes introduced in the higher level code that would occur in the experiment is the critical element of the predictive/descriptive model that is crucial for the design verification utility of the base class.

3.3 Daffodil-lib Hardware Class

Of the many concrete classes derived from the Daffodil-lib base class, one of them is uniquely privileged in that it controls the hardware directly, rather than running simulations. We call this the *hardware class.* The Daffodil-lib hardware class, like the simulation class, inherits its interface from the base class and keeps the descriptive behaviors. The predictive behaviors, such as the routing of signals, placement of voltages on the DACs/ADCs, or ReRAM response characteristics are now spontaneous reactions of the physical system rather than mathematical models.

The descriptive behaviors, such as setting the CMOS logic signals, programming the DACs/ADCs, and timing the operation of events must each invoke either individual GPIO lines or otherwise call device drivers from the device tree, which may either access individual components or orchestrate a complex series of hardware-timed actions. To avoid confusing syntax, Python's operator overloading features are used to expand the get attribute and set attribute actions of the base class to also simultaneously invoke the associated driver calls. Critically, none of the software interfaces between the hardware class and the base class are changed, allowing the higher level functions to operate with either the Daffodil-lib base or hardware class without modification. In addition, any programming errors that would lead to undesired behavior in the hardware class are predicted by the operations of the base class, allowing both model and simulation to co-verify one another.

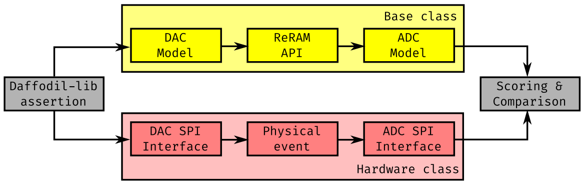


Figure 3: Polymorphism allows for commands to interface either to a simulation of the prototyping hardware or to be routed to the interface drivers.

4 DAFFODIL-APP, VERILATOR, AND NEURAL NETWORK RESULTS

The highest level code is composed of Daffodil-app, a separate library which takes all of the primitives in Daffodil-lib and maps it into dimensionless network operations. These include constructing network layers from multiple kernels, performing outer-product operations on arbitrarily sized layers constructed from kernel-level operations, and running inference/backpropagation operations. The library has predefined values for read and write voltages; algorithmic hyperparameters are the only user-tunable values required. In addition to providing a consistent interface between the simulation and hardware classes that renders them indistinguishable, the app also uses pybind11 to produce a consistent Python interface between C-code, C++ libraries generated from verilated SystemVerilog

Hoskins, Brian; Fream, Mitchell; Daniels, Matthew; Goodwill, Jonathan; Madhavan, Advait; McClelland, Jabez J.; Yousuf, Osama; Adam, Gina; Ma, Wen; Liu, Muqing ; Madsen, Rasmus; Lueker-Boden, Martin. "A System for Validating Resistive Neural Network Prototypes." Presented at International Conference on Neuromorphic Systems 2021, Knoxville, TN, US. July 27, 2021 - July 29, 2021.

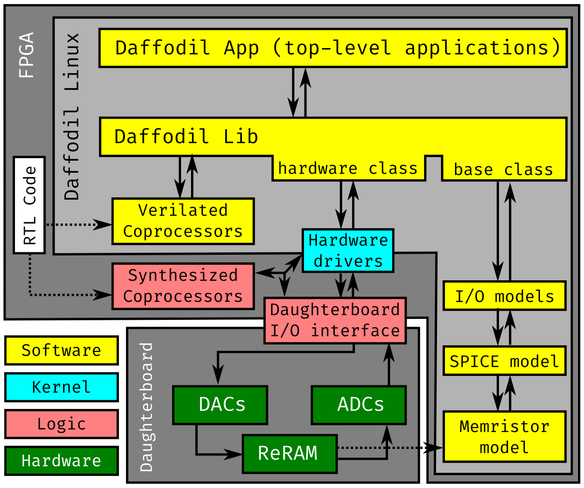


Figure 4: Depiction of the verification framework. The framework can be run from a simulation-only host, run within the FPGA CPU on the embedded OS, be partially synthesized on the FPGA, or interfaced into the physical hardware for the experiment. Solid arrows indicate dataflow; dashed arrows represent the derivation source/destination.

(which model the activity of our own custom compute accelerator or coprocessor modules), and compiled application-level drivers on the FPGA. Consequently, in addition to modeling/operating mixed signal computing, the Daffodil-app enables simulation of a hardware coprocessor from Python script models to a synthesized ASIC model. Using the Daffodil-app, we are able to engage with predictive modeling of network training on our hardware platform. In one model, we used 28 of the 32 available kernels to create a 324 × 50 × 10 network to study a reduced-MNIST problem. Using an ideal ReRAM model with conductance from 1 µS to 100 µS, we show that an operationally exact hardware model of our system can train to acceptable accuracy. In addition to running basic mini-batch gradient descent, we simulate the operation of coprocessors implementing reduced-rank, stochastic training [15, 16].

5 DISCUSSION AND FUTURE WORK

Constructing a design verification platform requires not only building a model of the hardware system, but also a model which is operationally faithful across multiple layers of abstraction. Disciplined end-to-end isomorphism is critical to the tracking of errors and side effects which otherwise might be untraceable when comparing a hardware model to an entirely external reference model.

Daffodil-lib does this through every layer of abstraction. The code models coprocessor operations that describe signals on the board. The ReRAM API includes only details of time and voltage, and does not presuppose any mode of operation in which any of its side effects may be absent. By building and exploiting polymorphism between simulation-control and hardware-control codebases, experimental errors are avoided. Imbuing hardware-control code with syntax identical to that of its simulation not only establishes this link for catching errors, but also reduces the cognitive load of moving from the modeling phase to implementation.

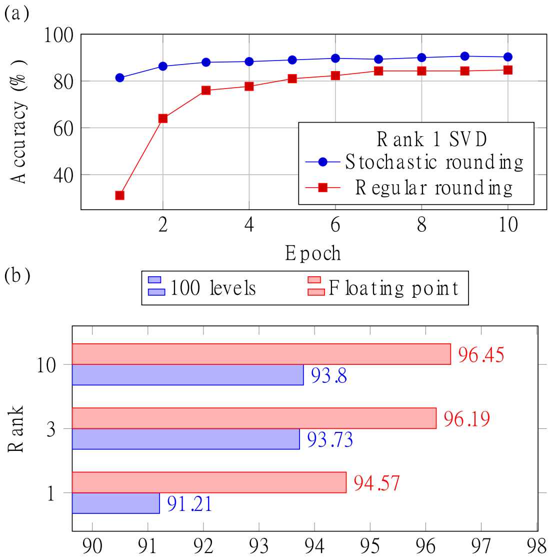


Figure 5: Plots of the simulated training on MNIST. Each point represents the test set accuracy for a single training. The networks were trained using a low rank training algorithm using both round nearest and stochastic rounding. a) Epochal performance of the training. b) Performance as a function of increasing the number of ranks.

This system brings many advantages even though simulations are slower than they could otherwise be. Modelling the ReRAM interactions as mixed-signal quantities in this co-simulation environment allows the system to debug errors in the hardware drive routines, or in the application loops. In trouble-shooting bad results, such mistakes can be distinguished from, higher-level problems such as poorly optimized hyperparameters.

The strong organizational correspondence between the hardware architecture and the python code is meant to facilitate a staged design process. Once an algorithm has been validated in simulation and with real devices, the whole design can be streamlined. Individual methods can be replaced with FPGA hardware co-processors with the end goal of completely realizing the end-to-end training data-path in RTL hardware. This top-down design approach can be repeated many times, and adapted to handle idiosyncracies introduced by adjustments in media processing, choice of material, or even experiments using a different memory technology.

6 CONCLUSION

In the context of the Daffodil-lib and associated components, we have discussed a developed operation exact design for verification framework for a resistive neural network prototype. The concepts proposed in it, based on decades of research into design verification and project management, from interfaces, to descriptive modeling, to synaptic polymorphism, can be used to thoughtfully grow a neuromorphic concept from a model to a working experiment.

ACKNOWLEDGMENTS

This work has been partially supported by the DARPA / ONR grant No. N00014-20-1-2031 and GWU University Facilitating Fund. AM acknowledges support under the NIST-UMD Cooperative Research Agreement Award No. 70NANB14H209, through the University of Maryland. Authors acknowledge Sonia Buckley and Mark Stiles for their helpful review of the manuscript.

REFERENCES

[1] Syed Ahmed Aamir, Paul Müller, Gerd Kiene, Laura Kriener, Yannik Stradmann, Andreas Grübl, Johannes Schemmel, and Karlheinz Meier. 2018. A mixed-signal structured adex neuron for accelerated neuromorphic cores. *IEEE transactions on biomedical circuits and systems* 12, 5 (2018), 1027–1037.

[2] Sapan Agarwal, Robin B Jacobs Gedrim, Alexander H Hsia, David R Hughart, Elliot J Fuller, A Alec Talin, Conrad D James, Steven J Plimpton, and Matthew J Marinella. 2017. Achieving ideal accuracies in analog neuromorphic computing using periodic carry. In *2017 Symposium on VLSI Technology*. IEEE, T174–T175.

[3] Christiaan Baaij, Matthijs Kooijman, Jan Kuper, Arjan Boeijink, and Marco Gerards. 2010. Cλash: Structural descriptions of synchronous hardware using haskell. In *2010 13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools*. IEEE, 714–721.

[4] Tobias Bjerregaard and Shankar Mahadevan. 2006. A survey of research and practices of network-on-chip. *ACM Computing Surveys (CSUR)* 38, 1 (2006), 1–es.

[5] Geoffrey W Burr, Robert M Shelby, Severin Sidler, Carmelo Di Nolfo, Junwoo Jang, Irem Boybat, Rohit S Shenoy, Pritish Narayanan, Kumar Virwani, Emanuele U Giacometti, et al. 2015. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices* 62, 11 (2015), 3498–3507.

[6] Pai-Yu Chen, Xiaochen Peng, and Shimeng Yu. 2018. NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 37, 12 (2018), 3067–3080.

[7] Wen Chen, Sandip Ray, Jayanta Bhadra, Magdy Abadir, and Li-C Wang. 2017. Challenges and trends in modern SoC design verification. *IEEE Design & Test* 34, 5 (2017), 7–22.

[8] John Clow, Georgios Tzimpragos, Deeksha Dangwal, Sammy Guo, Joseph McMahan, and Timothy Sherwood. 2017. A pythonic approach for rapid hardware prototyping and instrumentation. In *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 1–7.

[9] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. 2018. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro* 38, 1 (2018), 82–99.

[10] Laura Fick, David Blaauw, Dennis Sylvester, Skylar Skrzyniarz, M Parikh, and David Fick. 2017. Analog in-memory subthreshold deep neural network accelerator. In *2017 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 1–4.

[11] Tom Fitzpatrick. 2004. SystemVerilog for VHDL users. In *Proceedings Design, Automation and Test in Europe Conference and Exhibition*, Vol. 2. IEEE, 1334–1339.

[12] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, and Design Patterns. 1995. Elements of Reusable Object-Oriented Software. *Design Patterns. massachusetts: Addison-Wesley Publishing Company* (1995).

[13] Per Haglund, Oskar Mencer, Wayne Luk, and Benjamin Tai. 2003. Hardware design with a scripting language. In *International Conference on Field Programmable Logic and Applications*. Springer, 1040–1043.

[14] Nicholas C Harris, Jacques Carolan, Darius Bunandar, Mihika Prabhu, Michael Hochberg, Tom Baehr-Jones, Michael L Fanto, A Matthew Smith, Christopher C Tison, Paul M Alsing, et al. 2018. Linear programmable nanophotonic processors. *Optica* 5, 12 (2018), 1623–1631.

[15] Brian D Hoskins, Matthew W Daniels, Siyuan Huang, Advait Madhavan, Gina C Adam, Nikolai Zhitenev, Jabez J McClelland, and Mark D Stiles. 2019. Streaming batch eigenupdates for hardware neural networks. *Frontiers in neuroscience* 13 (2019), 793.

[16] Siyuan Huang, Brian D Hoskins, Matthew W Daniels, Mark D Stiles, and Gina C Adam. 2020. Memory-efficient training with streaming dimensionality reduction. *arXiv preprint arXiv:2004.12041* (2020).

[17] Shunning Jiang, Peitian Pan, Yanghui Ou, and Christopher Batten. 2020. PyMTL3: a Python framework for open-source hardware modeling, generation, simulation, and verification. *IEEE Micro* 40, 4 (2020), 58–66.

[18] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. 2017. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*. 1–12.

[19] Malte J Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. 2021. A flexible and fast PyTorch toolkit for simulating training and inference on analog crossbar arrays. *arXiv preprint arXiv:2104.02184* (2021).

[20] Rindert Schutten and Tom Fitzpatrick. 2003. Design for verification.

[21] John Sweeney. 2020. Techniques for performing design verification of an optical processor. https://medium.com/lightmatter/techniques-for-performing-design-verification-of-a-mixed-signal-digital-analog-and-optical-d0068c9ff868

[22] Cheng Tan, Yanghui Ou, Shunning Jiang, Peitian Pan, Christopher Torng, Shady Agwa, and Christopher Batten. 2019. PyOCN: A unified framework for modeling, testing, and evaluating on-chip networks. In *2019 IEEE 37th International Conference on Computer Design (ICCD)*. IEEE, 437–445.

Requirement elicitation for adaptive standards development

Marion Toussaint [*,**] Sylvere Krima [***]
Allison Barnard Feeney [****] Herve Panetto [**]

[*] *Associate, NIST, 100 Bureau Drive, Gaithersburg, MD, 20899, USA,*
(e-mail: marion.toussaint@nist.gov)
[**] *Université de Lorraine, CNRS, CRAN, 54000 Nancy, France*
[***] *Engisis LLC, 10411 Motor City Dr Ste 750, Bethesda, MD,*
20817-1289, USA, (email: sylvere.krima@engisis.com)
[****] *NIST, 100 Bureau Drive, Gaithersburg, MD, 20899, USA*

Abstract: The recent digitization of manufacturing, also referred to as the fourth industrial revolution (or Industry 4.0), heavily relies on information standards for the exchange and integration of digital data across manufacturers and their partners. Standards are complex projects with a long lifecycle, often developed in a predictive environment, which no longer aligned with a fast paced industry. Adaptive environments have been proven to be an answer to this problem, but most project requirements management solutions have not evolved to support this shift. In this paper, we discuss the new challenges brought by a shift to adaptive environments, and introduce a solution to offer better requirement elicitation during standards development management with increased traceability and visibility.

Keywords: information standards, requirement elicitation, requirement model, agile, adaptive development

1. INTRODUCTION

Information standards are powerful tools for innovation and productivity in many domains (Gallaher et al., 2004; Blind, 2009; Guasch et al., 2007). When it comes to manufacturing, information standards are seen as a key enabler to the digitization of the manufacturing industry (Fischer et al., 2015; Helu et al., 2017; Hedberg et al., 2016). But standards development is relatively complex. It is a long process that includes many stakeholders, from different organizations, geographically dispersed, and on a volunteer basis, making their contribution and participation irregular. The human resources available depend on the experts' schedules and their organizations' needs, which makes it difficult to have continuity and consistency in standards development. Moreover, because their participation is irregular, it becomes difficult for every member of the standard development process to keep up with the past and current activities of the development. To summarize, the standard development process is long, irregular, and difficult to plan.

Information standards are mainly developed using the predictive or waterfall model for project management. The waterfall model is a sequential development model, in which each phase of the development must be completed in order to proceed to the next phase (Balaji and Murugaiyan, 2012). In predictive models such as the waterfall model, the project requirements and deliverables are defined at the beginning of the project (PMI, 2017), and if some requirements need to be changed or added, they will most likely not be implemented in the current development iteration (Balaji and Murugaiyan, 2012). According to ISO, standards development iterations last between 18 and 48 months (ISO, 2017), which means that: i) in the best-case scenario, new requirements will be addressed up to 18 months after being identified, ii) in the worst-case scenario, new requirements will be addressed up to 48 months after being identified. The current length and management of standards development iterations are not necessarily adapted to the needs of the industry. In the industry, strong market competition leads to a short product life cycle (Sapp et al., 2021) and requirements change often and faster than the development of standard iterations. Besides, during the long development iterations, new technologies, processes, and needs are developed, potentially making the published standards at odds with the industrial reality.

An alternative to the waterfall model for project management is to use an adaptive method (also known as Agile) (Shameem et al., 2018; Thummadi et al., 2011). Unlike waterfall, Agile follows an iterative and incremental approach. This model consists of short iterative release cycles, which requires more transparency and visibility, stakeholders to be more involved and notified more regularly of the progress of the project (Edeki, 2015). Agile is mainly used to achieve high quality projects in short periods, better collaboration between all stakeholders, and reduced time to market (Kumar and Bhatia, 2012). Agile can be implemented through many frameworks such as Scrum, XP, or Lean Kanban (Stellman and Greene, 2013). Most Agile frameworks are designed for small development teams. However, standards development often requires several (large) teams to work together. The SAFe Framework

Toussaint, Marion; Krima, Sylvere; Barnard Feeney, Allison; Panetto, Herve. "Requirement elicitation for adaptive standards development."
Presented at 17th IFAC Symposium on Information Control Problems in Manufacturing (INCOM 2021). June 07, 2021 - June 09, 2021.

by Scaled Agile (Agile, 2019) is a viable alternative for standards development due to its support of large, multidisciplinary, and distributed teams. SAFe methods can bring benefits to the development teams of model-based standards, as demonstrated in (Sapp et al., 2021). The report (Sapp et al., 2021) demonstrates the benefits of adopting an agile framework and tool-chain in support of the standards development processes, illustrated with ISO 10303 (Pratt, 2005), a large and complex standard widely used in the manufacturing industry.

While becoming agile can bring many benefits to standards development processes, it also creates some challenges. Because requirements are defined, analyzed, and processed on a more frequent basis, properly capturing, tracking, and managing them becomes crucial, more time consuming, and more complex. While the different phases of requirements engineering present many challenges (Besrour et al., 2016), most of the literature focuses on conflict management (Shah and Patel, 2014; Davis et al., 2006; In and Roy, 2001; Gambo et al., 2015; Decker et al., 2007) and predictive environment, leaving out requirements elicitation, a key step to their management (PMI, 2017). In this paper, we present a solution for requirements elicitation in an adaptive information standards development environment that enables maximal transparency, traceability, and visibility of the project activities.

2. RELATED WORK

Requirements elicitation is a critical step in project development and is sometimes considered a complex process (Sharma and Pandey, 2013; Fernandes et al., 2012). Many methods have been developed to capture and represent requirements to facilitate this process. In this section, we present some of the most representative methods of the state of the art.

The Requirements Interchange Format, also known as ReqIF, is a standardized XML-based format aimed to support the exchange of requirements, as well as their associated metadata, between different management tools (Adedjouma et al., 2011). This nonproprietary format satisfies the industry needs for exchanging requirements data between different companies without having to share the same management tools (OMG, 2016). ReqIF is supported by almost all existing requirements management tools and is widely used. A ReqIF model is composed of specification objects, called SpecObject, to represent the requirements. These objects contain multiple user-defined attributes and the relationships between different objects are also represented.

The Systems Modeling Language (SysML) is a modeling language for the specification, analysis, design, and verification of a broad range of complex systems (SysML, 2019). SysML includes different types of diagrams, one of which is used to represent text-based requirements. This diagram can effectively capture functional as well as non-functional requirements. This SysML diagram captures the relationship between requirements, system models, and test cases (Hause, 2006). A SysML requirement is mainly defined by a unique identifier and a text-based definition, but other properties can be included such as priority, status, or verification status (Roques, 2015). An example of a SysML requirement diagram for an Automobile System is shown in Figure 1. This example illustrates the capture of text-based requirements and the relations between them.

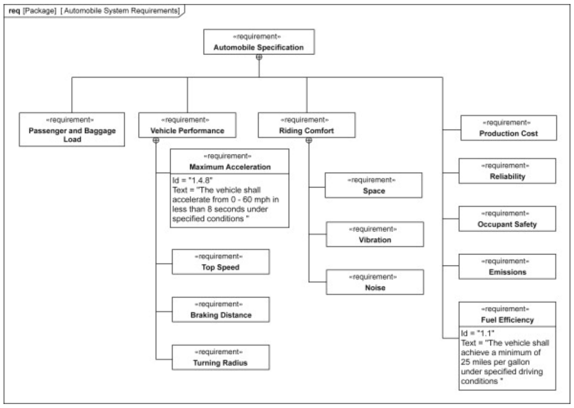

req [Package] [Automobile System Requirements]
«requirement» Automobile Specification
«requirement» Passenger and Baggage Load
«requirement» Vehicle Performance
«requirement» Riding Comfort
«requirement» Production Cost
«requirement» Maximum Acceleration
Id = "1.4.8"
Text = "The vehicle shall accelerate from 0 - 60 mph in less than 8 seconds under specified conditions "
«requirement» Reliability
«requirement» Space
«requirement» Vibration
«requirement» Occupant Safety
«requirement» Top Speed
«requirement» Emissions
«requirement» Noise
«requirement» Braking Distance
«requirement» Fuel Efficiency
Id = "1.1"
Text = "The vehicle shall achieve a minimum of 25 miles per gallon under specified driving conditions "
«requirement» Turning Radius

Fig. 1. Example of a SysML requirements diagram for the Automobile System Specification. (S. Friedenthal and Steiner, 2015)

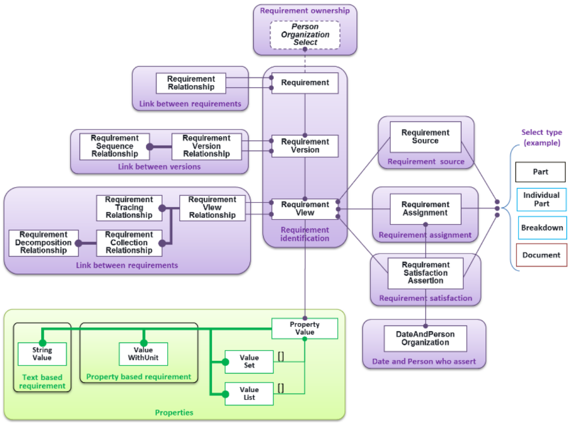

Requirement ownership
Person Organization Select
Requirement Relationship
Link between requirements
Requirement
Requirement Sequence Relationship
Requirement Version Relationship
Link between versions
Requirement Version
Requirement Source
Requirement source
Select type (example)
Part
Individual Part
Breakdown
Document
Requirement Tracing Relationship
Requirement View Relationship
Requirement View
Requirement identification
Requirement Assignment
Requirement assignment
Requirement Decomposition Relationship
Requirement Collection Relationship
Link between requirements
Requirement Satisfaction Assertion
Requirement satisfaction
Property Value
DateAndPerson Organization
Date and Person who assert
String Value
Value WithUnit
Value Set
[]
Value List
[]
Text based requirement
Property based requirement
Properties

Fig. 2. STEP AP 242 Requirements model. (AP242, 2013)

ISO 10303, also known as the Standard for the Exchange of Product data (STEP), is a widely adopted standard that aims to provide a complete and unambiguous description of manufacturing products, usable throughout their life cycle, regardless of the IT support used. STEP enables manufacturing applications to exchange and share data regardless of the specificities of the different systems that exchange and share information. STEP AP242 (ISO, 2020) includes a model for requirements management, as illustrated in Figure 2. This model focuses principally on capturing requirements definitions, as well as their relationships, validation criteria, and source (AP242, 2013).

These models focus mainly on a limited and textual definition of the requirements, and the relationships between them. These models were designed to support the representation and exchange of requirements and not meant to support advanced requirements management and planning activities (PMI, 2017) in an adaptive environment. As discussed previously, in transitioning from a waterfall to an adaptive project management approach, requirements management takes a different and more important role, and requires more information and more visibility, to be executed properly (Institute, 2017). Besides, working with distributed teams and volunteer human resources requires better traceability and visibility of both decisions and contributions. An ideal solution should focus on defining the information to capture in order to overcome the following challenges (Sapp et al., 2021):

- (N1) It should ensure the traceability (and visibility) of resources: 1) to associate people to work items and tasks to which they contribute or supervise (i.e., their level of engagement), 2) to link people to the meetings they attend, and 3) to link work items to meetings during which they are discussed or to follow the progress of the work items;
- (N2) It should ensure the traceability (and visibility) of decisions: 1) by linking people to the decision they make, 2) by linking decisions to the meetings during which they are taken, and 3) by following the evolution of decisions regarding work items and requirements as the meetings progress;
- (N3) It should ensure requirements definition management by linking requirements to their source and a context in order to be able to validate them with their owner.

3. PRESENTATION OF THE MODEL

None of the models presented in the previous section captures enough information to overcome the challenges we presented because they lack a (precise) definition of 1) the requirements' context and source, 2) the requirement ownership information, 3) the associated resource(s) management, 4) the associated work breakdown, 5) key decisions, and 6) tracking information between requirements and associated deliverables. Consequently, we developed our own requirements elicitation model. This model can be used on its own or as an extension to the previously introduced solutions (i.e., ReqIF, SysML, and AP242). Figure 3 illustrates this elicitation model we propose. Our model is a UML Class diagram and is composed of six classes (Meeting, Requirement, Work Item, Task, Member, and Decision) in response to our three main needs: requirements definition management and traceability, decisions traceability, and commitment and resources traceability. This model supports our needs as shown in Table 1.

The objective of the proposed model is to keep track of decisions and requirements throughout the development of a project, which is crucial for long projects involving different distributed stakeholders. This model can be used to understand the level of engagement of the different stakeholders and thus be able to more easily manage the available resources. Moreover, this model allows us to record all decisions taken during meetings, making that information available to people who could not attend. It also allows us to ensure the traceability of requirements from their creation to their implementation in order to validate them directly with the people who had expressed these needs in the first place.

3.1 Meeting Class

The *Meeting* class represents the different properties that characterize a meeting and contains several properties to keep track of all the details of the meeting:

- The *id* property uniquely identifies the meeting;
- The *event* property defines the event during which the meeting took place, but this property is not mandatory because meetings are not necessarily included in a particular event;
- The *place*, *date* and *time* properties correspond respectively to the place, the date, and the time at which the meeting was held;
- The *type* property defines if the meeting is in person or virtual. If the meeting is virtual, the *place* property is left empty;
- The *subject* property defines the main subject of the meeting from a high-level point of view, while the *topic* property represents the agenda of the meeting which is the list of items that need to be discussed during the meeting;
- The *goal* property represents the objective of the meeting;
- The *future points* property represents topics that will have to be discussed during the next meeting.

The *Meeting* class has relationships with some of the other classes. Firstly, the Meeting class is linked to the *Member* class: several people can attend a meeting. Secondly, there are two relationships between the *Meeting* and *Requirement* classes: requirements are created during a meeting, and requirements can be discussed and updated during one or more meetings. Thirdly, the *Meeting* class is linked to the *Work item* class: work items can be planned during one or more meetings.

3.2 Member Class

The *Member* class represents the characteristics of a stakeholder. A member can be anyone involved in the project such as a manager, a developer, or a client. A member contains two properties, the name of the person and the organization to which that person belongs. The *Member* class is linked to the following classes:

- To the *Meeting* class: several persons can present one or more topics during a meeting, but the presenter(s) must be one of the meeting's attendees;
- To the *Decision* class: every member can express an opinion about requirements or a work item;
- To the *Requirement* class: a requirement has one or more authors;
- To the *Work item* class across two different relationships: one relation represents the work item's supervisor while the other one represents the team members working on it. We considered that a supervisor contributes to the work item that he supervises and that it is not necessary to create multiple relation links between a member and the same work item. Several work items can have the same supervisor and one or more team members working on it;
- To the *Task* class: one or more persons can contribute to some tasks.

Table 1. List of the UML classes from our elicitation model and the needs they meet

Needs	**Model Classes**
Requirements definition management and traceability (N3)	Member, Meeting and Requirements classes
Decisions traceability (N2)	Decision, Member and Meeting classes
Commitment and resources traceability (N1)	Requirement, Work Item, Task, Meeting and Member classes

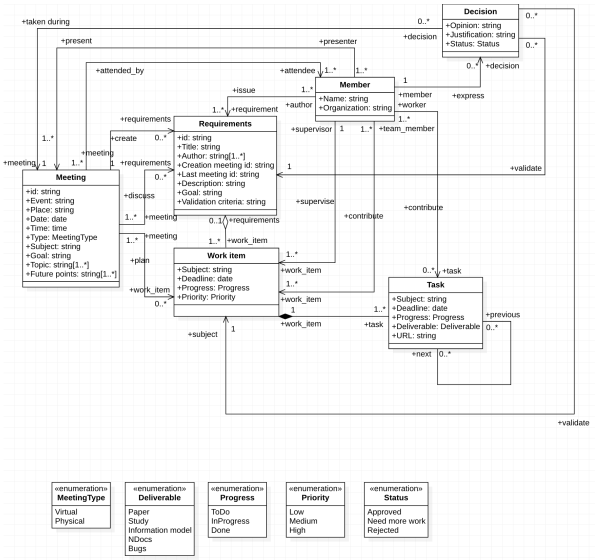


Fig. 3. Requirements elicitation model (UML Class Diagram)

3.3 Decision Class

During a meeting, each attendee is free to express an opinion on the different topics discussed. The objective of the *Decision* class is to keep track of the decisions expressed during a meeting. This class contains three properties to characterize these decisions. An attendee's opinion on a particular requirement or work item and its justification are respectively represented by the *opinion* and *justification* properties. The last property defines the status of the decision and this property can currently only take one of the following three values: *Approved*, *Need more word*, or *Rejected*.

The *Decision* class has relationships with three other classes: the *Meeting* class, the *Requirement* class, and the *Work Item* class. A decision is linked to the meeting during which this decision was taken, and to the requirement or work item to which this decision relates. Several decisions can be taken during the same meeting and in the same way, several decisions can be associated with one requirement or work item.

3.4 Requirement Class

The *Requirement* class represents the different properties that characterize a requirement. Thus, a requirement is composed of several properties. Firstly, a requirement is identified by a unique id. A requirement is also characterized by a title and an author. Besides, the *Requirement* class contains the id of the meeting during which the requirement was created and the id of the last meeting during which the requirement was updated. When a requirement is created, the *creation meeting id* and *last meeting id* properties are identical. Finally, a requirement is also characterized by its description, its goal, and its validation criteria.

3.5 Work Item Class

A project is made up of several work items that represent the different steps to reach the goal of the project. Work items help to plan and manage projects. In our case, one or more work items can be associated with a requirement. The work item is characterized by a high-level subject, a deadline, its progress, and a priority status. A work item's progress can either be *To Do*, *In Progress*, or *Done*, while its priority can either be *Low*, *Medium*, or *High*. The progress of a work item depends on the progress of the tasks that compose it: indeed, a work item cannot be marked as *Done* if all its tasks are not also marked as *Done*.

3.6 Task Class

A work item is composed of one or more tasks. The *Task* class can represent user stories, tasks, bugs, or issues. A task is characterized by five properties: a subject which defines the goal of the task, a deadline, its progress, whose value is included among *To Do*, *In Progress* or *Done*; the type of deliverable returned, and the url associated with the commit corresponding to the task's deliverable. Currently, the deliverable of a task can be a *Paper*, an *Information model*, some *NDocs*, or a *Bug*. When the task has not yet started, the url property is left empty. In some cases, the different tasks included in a work item need to respect a certain order, represented by the relationship between the tasks.

4. DEMONSTRATION OF THE MODEL

As presented in the previous section, this model allows us to keep track of and share information regarding project meetings, decisions, requirements, and work items. This model also keeps all the stakeholders informed about the progress of the project and the different topics discussed during the different project meetings. Moreover, this model supports tracking traceability of the requirements by keeping track of their creation and their authors, as well as any changes that may have been made and the progress of the associated work items. This information is key to better understand and follow the evolution of the requirements over time and to validate them directly with the people who expressed them in the first place.

The Figure 4 illustrates a virtual technical meeting attended by five stakeholders. The main objective of this project meeting is the project management methodology with the transition to Agile. During this meeting, the 45e65d requirement on Agile as a project management method was updated and two work items were discussed. First, the “Agile” work item, associated with the 45e65d requirement that is composed of four tasks. The first task has been completed, the second is still in progress, while the other two tasks were just added during this meeting. As new tasks, they do not have contributors yet, they don’t have an associated repository, and their deliverables are not yet defined. Second, the work item concerning the review of the issues log is still in progress. During the meeting, the bugs were reviewed and their resolution was planned.

Due to its design, our solution can be used on its own or integrated to an existing model. For instance, to integrate this solution to ReqIF, one would replace the *SpecObject* element of the ReqIF model with our *Requirement* class and/or replace the *System Component* element of ReqIF with our *Work Item* class

5. CONCLUSION

In this paper, we discussed the role of information standards, a key enabler to the digitization of manufacturing, and the inefficiencies in their current development process that often relies on a predictive management approach. As an alternative, we highlighted how an adaptive approach that can overcome some of these inefficiencies also comes with its own challenges, notably an increased need for traceability and visibility of requirements, their elicitation, their management, and their implementation. As this need had not been addressed, we developed and proposed a new requirement elicitation model that serves as a foundation for providing requirements and decision traceability and visibility, by recording and leveraging project meetings in a formal way.

This information model, unlike existing traditional solutions, has been designed for adaptive management and captures all the information essential to properly implement an agile approach. Due to its simplicity, the model can easily be used on its own or as an extension to an existing requirements management solution. As an extension, it can extend and enrich traditional solutions that were not initially designed for adaptive management.

Once this model has been properly implemented and instantiated (i.e., data has been collected), the next challenge is to display the captured information in a meaningful way to the different project stakeholders. Because visibility is key to agile management, project information is often shared graphically in an information radiator that is easily accessible and understandable by all. Our future work will focus on identifying the most appropriate visualization techniques to display the information we are now able to collect.

REFERENCES

Adedjouma, M., Dubois, H., and Terrier, F. (2011). Requirements exchange: From specification documents to models. In *2011 16th IEEE International Conference on Engineering of Complex Computer Systems*. IEEE.

Agile, S. (2019). Safe 5.0 framework. URL `https://www.scaledagileframework.com/`. Accessed: 2020-10-10 [Online].

AP242, S. (2013). Requirement management interoperability. URL `http://www.ap242.org/requirement-interoperability`. Accessed: 2020-10-13 [Online].

Balaji, S. and Murugaiyan, D.M.S. (2012). Waterfall vs v-model vs agile: A comparative study on sdlc. *International Journal of Information Technology and Business Management*, 2(1).

Besrour, S., Rahim, L.B.A., and Dominic, P.D.D. (2016). A quantitative study to identify critical requirement engineering challenges in the context of small and medium software enterprise. In *2016 3rd International Conference On Computer And Information Sciences (ICCOINS)*. IEEE.

Blind, K. (2009). Standardisation as a catalyst for innovation. *ERIM Report Series*.

Davis, C.J., Fuller, R.M., Tremblay, M.C., and Berndt, D.J. (2006). Communication challenges in requirements elicitation and the use of the repertory grid technique. *Journal of Computer Information Systems*, 46(5).

Decker, B., E. Ras, J.R., Jaubert, P., and Rieth, M. (2007). Wiki-based stakeholder participation in requirements engineering. *IEEE Software*, 24(2).

Edeki, C. (2015). Agile software development methodology. *European Journal of Mathematics and Computer Science*, 2(1).

Fernandes, J., Duarte, D., Ribeiro, C., Farinha, C., Pereira, J.M., and da Silva, M.M. (2012). ithink : A game-based approach towards improving collaboration and participation in requirement elicitation. *Procedia Computer Science*, 15.

Fischer, K., Rosche, P., Trainer, A., Feeney, A.B., and Hedberg, T.D. (2015). Investigating the impact of standards-based interoperability for design to manufacturing and quality in the supply chain.

Gallaher, M.P., O’Connor, A.C., John L. Dettbarn, J., and Gilday, L.T. (2004). Cost analysis of inadequate

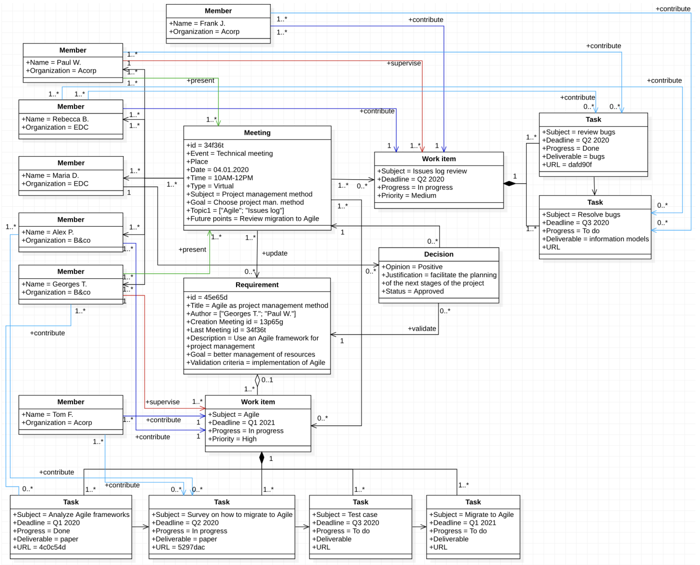


Fig. 4. Simple example of our Requirement Elicitation model (UML Object Diagram)

interoperability in the u.s. capital facilities industry. *National Institute of Standards and Technology (NIST)*.

Gambo, I.P., Soriyan, A.H., and Ikono, R.N. (2015). A proposed process model for requirements engineering using delphi techniques for prioritization. *International Journal of Information Technology and Computer Science*.

Guasch, J., Racine, J.L., Sánchez, I., and Diop, M. (2007). *Quality Systems and Standards for a Competitive Edge*. The World Bank.

Hause, M. (2006). The sysml modelling language. In *Fifth European Systems Engineering Conference*. INCOSE.

Hedberg, T.D., Feeney, A.B., Helu, M.M., and Camelio, J.A. (2016). Towards a lifecycle information framework and technology in manufacturing. *ASME Journal of Computing and Information Science in Engineering*.

Helu, M.M., Hedberg, T.D., and Feeney, A.B. (2017). Reference architecture to integrate heterogeneous manufacturing systems for the digital thread. *CIRP Journal of Manufacturing Science and Technology*.

In, H. and Roy, S. (2001). Visualization issues for software requirements negotiation. In *25th Annual International Computer Software and Applications Conference*. IEEE.

Institute, P.M. (2017). *Agile Practice Guide*. Project Management Institute.

ISO (2017). Target date planner. Accessed: 2020-10-10 [Online].

ISO (2020). Industrial automation systems and integration — product data representation and exchange — part 242: Application protocol: Managed model-based 3d engineering.

Kumar, G. and Bhatia, P.K. (2012). Impact of agile methodology on software development process. *International Journal of Computer Technology and Electronics Engineering*, 2(4).

OMG (2016). Requirements interchange format (reqif). URL `https://www.omg.org/reqif/`. Accessed: 2020-10-11 [Online].

PMI (2017). A guide to the project management body of knowledge (pmbok guide). Project Management Institute, 6 edition.

Pratt, M.J. (2005). Iso 10303, the step standard for product data exchange, and its plm capabilities. *International Journal Of Product Lifecycle Management*, 1(1).

Roques, P. (2015). How modeling can be useful to better define and trace requirements. *Requirements Engineering Magazine*. Accessed: 2020-10-13 [Online].

S. Friedenthal, A.M. and Steiner, R. (2015). *A Practical Guide to SysML, Third Edition*. Morgan Kaufmann.

Sapp, B., Harvey, M., Toussaint, M., Krima, S., Feeney, A.B., and Panetto, H. (2021). Agile for model-based-standards development. *NIST Advanced Manufacturing Series 100-40*.

Shah, T. and Patel, S.V. (2014). A review of requirement engineering issues and challenges in various software development methods. *International Journal of Computer Applications*, 99(15).

Shameem, M., Kumar, C., Chandra, B., and Khan, A. (2018). Impact of requirements volatility and flexible management on gsd project success: A study based on the dimensions of requirements volatility. *International Journal of Agile Systems and Management*.

Sharma, S. and Pandey, S.K. (2013). Revisiting requirements elicitation techniques. *International Journal of Computer Applications*, 75(12).

Stellman, A. and Greene, J. (2013). *Learning Agile: Understanding Scrum, XP, Lean, and Kanban*. O'Reilly Media.

SysML (2019). Sysml open source project. URL `https://sysml.org`. Accessed: 2020-10-11 [Online].

Thummadi, B., Shiv, O., and Lyytinen, K. (2011). Enacted routines in agile and waterfall processes. In *2011 Agile Conference*. IEEE.

Deterministic Tagging Technology for Device Authentication

Jungjoon Ahn
Semiconductor and Dimensional Metrology Division, Physical Measurement Laboratory
National Institute of Standards and Technology (NIST)
Gaithersburg, MD US
Jungjoon.ahn@nist.gov

Jihong Kim
Department of Electrical Engineering
Yeungnam University
Gyeongsan, Republic of Korea
jihongkim@yu.ac.kr

Joseph J. Kopanski
Semiconductor and Dimensional Metrology Division, Physical Measurement Laboratory
National Institute of Standards and Technology (NIST)
Gaithersburg, MD US
joseph.kopanski@nist.gov

Yaw S. Obeng
Semiconductor and Dimensional Metrology Division, Physical Measurement Laboratory
National Institute of Standards and Technology (NIST)
Gaithersburg, MD US
yaw.obeng@nist.gov

***Abstract*—This paper discusses the development of a rapid, large-scale integration of deterministic dopant placement technique for encoding information in physical structures at the nanoscale. The doped structures bestow a customizable radiofrequency (RF) electronic signature, which could be leveraged into a distinctive identification tag. This will allow any manufactured item (integrated circuit, pharmaceutical, etc.) to be uniquely authenticatable. Applications of this technology include enabling a secure Internet of Things (IoT) and eliminating counterfeit products.**

Keywords*—*Nanoelectronics, reliability, metrology, system security, probe assisted deterministic doping, PAD

I. Introduction

The Internet of Things (IoT) represents a new era of telecommunications made possible by the reduced cost of performance in electronic devices, the convergence of wireless technologies, advancements of analog systems (e.g., MEMS) and digital electronics (i.e. More-than-Moore technologies). How these devices connect to each other, and to humans, are changing how we work and live. Unfortunately, the weaknesses of the underlying networks have been exposed through the exploitation of hardware operation weaknesses. Without ample security measures, the ever-expanding sensor network could create massive vulnerabilities [1-2]. Hardware security based on a dynamic electronic tagging system that supports unique and encrypted identification is a prerequisite component of the security envelope.

The ability to tag an item by deterministically placing small clusters of dopants at the nanoscale reveals some interesting possibilities for device fabrication [3]. One promising application is the selective placement of dopants to tune the electrical properties of nano-channel MOSFETs [4-7]. Another possibility is the implantation of dopant clusters to modify the barrier height in metal-semiconductor junctions.

The probe assisted doping technique (PAD) provides the ability to rapidly produce customizable 2D-superlattices of p-n junctions on a semiconductor substrate by controlling the dopant concentration of each element (Fig. 1.) [8].

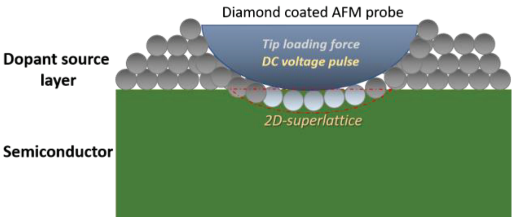


Fig. 1. Schematic representation of PAD.

PAD uses the tip-loading force and bias pulse of the probe in scanning probe microscopy (SPM) to achieve precise area control and subsequent verification imaging to implant dopants from a thin over-layer of source material. In contrast to other deterministic doping techniques, relatively large (20 x 20) 2-D superlattices can be readily formed with minimal impact to the semiconductor surface. The other advantages of PAD include (i) multiple patterns of doped semiconductor without photolithography and (ii) highly selective doping as compared to traditional invasive ion-implantation processes. While other deterministic doping processes (e.g., laser-enhance deposition and the single-ion on-demand technique) may have some advantages, they are more complex and require significant equipment investment [9]. The degrees of freedom in design and fabrication of PAD promotes a new class of low-cost identification tags for complex integrated semiconductor devices. Since photo-lithography process are not involved, each element of the array can be uniquely programmed with different structures and elemental configuration.

As an initial demonstration, clusters of aluminum (Al) atoms were deterministically doped into a wafer of n-type Si (100) to generate nanoscale counter-doped junctions within a

Ahn, Jungjoon; Kopanski, Joseph; Obeng, Yaw S.; Kim, Jihong. "Deterministic Tagging Technology for Device Authentication." Presented at IEEE International Conference on Integrated Circuit Design and Technology (ICICDT), Dresden, DE. September 15, 2021 - September 17, 2021.

few nanometers from Si-air interface. The resulting local electrical potential changes, reported as the contact potential difference (CPD), were verified with a scanning Kelvin probe microscope (SKPM). The electrical activation of nanojunctions was achieved after thermal annealing. Thermal activation, however, also promoted the diffusion of the dopants that results in an expansion of the deterministically doped sites [8].

II. EXPERIMENTAL

We adopted and modified the nano-indentation mode of the atomic force microscopy (AFM). A diamond coated tip (tip radius ~ 50 nm) on a stiff cantilever (spring constant ~ 60 N/m) rapidly created a 20 x 20 array of Al-injected p-n junctions over a 10 µm x 10 µm area of n-type Si substrates (1 Ω·cm to 10 Ω·cm). A thin Al dopant layer of 10 nm was deposited over the Si substrates using e-beam evaporator. The modified nano-indentation of each doping site was conducted by applying (i) identical trigger-threshold voltages (i.e., tip loading force) between 6 V to 8 V and (ii) a voltage pulse of 8 V that were adopted from the previously PAD experiments [8]. The x-y radius of each resulting injection site varied from 65 nm to 100 nm.

After PAD, the Al layer was chemically removed using a selective etchant. The resulting Si substrate was annealed for 30 sec that lead to diffusion of Al dopants and an activation. For post-PAD process measurements, we used scanning Kelvin force microscopy (SKFM) to measure the surface potential between the tip coating and the surface.

III. RESULTS AND APPLICATIONS

We used PAD to create arrays of Al-Si junctions as a potential nanoscale equivalent to the two-dimensional Quick Response (QR) code.

Fig. 2 shows the resulting surface potential difference after PAD of Al into the Si substrate. The potential differences varied from -10 mV to -50 mV as compared to undoped areas of the Si substrate. The greater potential differences correlated with the tip loading force gradually increased from 6 V in the bottom row to 8V in the top row of Fig. 2. Post-PAD measurements are not limited to SKFM and could employ other scanning probe microscopy techniques, such as capacitance measurements with scanning microwave microscope (SMM).

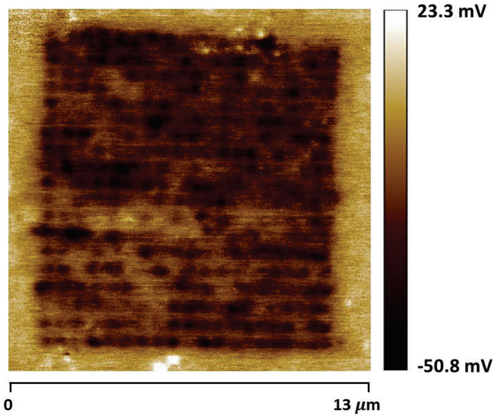


Fig. 2. SKFM image Al dopants injected and activated n-Si surface via PAD. A platinum (Pt) coated tip was used for SKFM

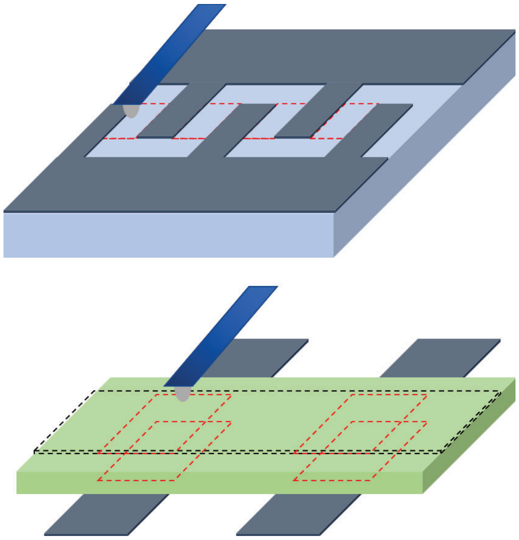


Fig. 3. Schematic representation of PAD for the device application (red dashed line). (Top) Semiconductor channel. (Bottom) Metal/semiconductor interface.

In theory, the modified surface potential changes should be directly dependent on the changes in the local crystallographic structure of the Si substrate [10] and the presence of dopants. However, we noticed some inconsistencies in the surface potential differences between areas doped under the same tip loading force, which we attributed to the non-uniformities in the Al overlayer.

The PAD arrays embedded within the silicon substrate have minimal topographic impact, which allows for direct processing on semiconductor layers or within layers deposited specifically for this process, such as a thin organic semiconductor or metal layer (Fig. 3). Unlike other substrate modifying techniques, such as through silicon visa (TSV), PAD is not expected to alter the stress in the substrate nor impact device performance [11]. Furthermore, the substrate doping technique is not limited by device geometry or dimensions [12].

PAD has the potential of storing large amounts of dense data with extensive encryption and sophisticated error correction algorithms. While optically readable codes have two levels (black and white), electromagnetic codes conceivably could have multiple levels of charge, potential, or magnetic moments to expand the data density beyond the simple bit. Arrays of such information could become an intrinsic, un-removable, un-alterable, un-forgeable and encrypted part of an integrated circuit. Thus, we propose PAD as a device tagging technique to enable tamper-proof authentication certificates for devices on the internet [13-14].

Ahn, Jungjoon; Kopanski, Joseph; Obeng, Yaw S.; Kim, Jihong. "Deterministic Tagging Technology for Device Authentication." Presented at IEEE International Conference on Integrated Circuit Design and Technology (ICICDT), Dresden, DE. September 15, 2021 - September 17, 2021.

IV. CONCLUSIONS AND FUTURE WORK

The ideal tag would be rapid and cheap to apply and read, un-removable, un-alterable, un-forgeable and encrypted. Our proposed new taggant system, via PAD, could meet these criteria with minimal operational cost. Each system would contain an intrinsic electronic component, traceable to the manufacturer and readable at any inspection point. The PAD technique is flexible enough to accommodate application specific customization. Future work will entail the optimization and automation of massively parallel deterministic doping [15], selection of suitable electromagnetic (EM) active dopants with high-reporting cross-sections and the development of the appropriate EM detectors/sensors.

REFERENCES

[1] Y. S. Obeng, C. Nolan and D. Brown, "Hardware security through chain assurance," 2016 Design, Automation & Test in Europe Conference & Exhibition, pp. 1535-1537, 2016.

[2] Y. S. Obeng, "Hardware Security to Mitigate Threats to Networked More-Than-Moore Sensors", 2016 ECS Trans., vol. 72, pp. 113, 2016.

[3] M. Hori, T. Shinada, Y. Ono, A. Komatsubara, K. Kumagai, T. Tanii, et al., "Impact of a few dopant positions controlled by deterministic single-ion doping on the transconductance of field-effect transistors," Applied Physics Letters, vol. 99, p. 062103, 2011

[4] J. Meijer, T. Vogel, B. Burchard, I. W. Rangelow, L. Bischoff, J. Wrachtrup*, et al.*, "Concept of deterministic single ion doping with sub-nm spatial resolution," Applied Physics A, vol. 83, pp. 321-327, 2006.

[5] R. W. Keyes, "Physical limits of silicon transistors and circuits," Reports on Progress in Physics*,* vol. 68, p. 2701, 2005.

[6] T. Shinada, S. Okamoto, T. Kobayashi, and I. Ohdomari, "Enhancing semiconductor device performance using ordered dopant arrays," Nature, vol. 437, pp. 1128-1131, 2005.

[7] H. Sellier, G. P. Lansbergen, J. Caro, S. Rogge, N. Collaert, I. Ferain*, et al.*, "Transport Spectroscopy of a Single Dopant in a Gated Silicon Nanowire," Physical Review Letters*,* vol. 97, p. 206805, 2006.

[8] J. -J. Ahn, S. D. Solares, L. You, H. Noh, J. Kopanski, and Y. Obeng , "Probe assisted localized doping of aluminum into silicon substrates", Journal of Applied Physics vol. 125, 075706, 2019.

[9] E. Prati and T. Shinada, Single-Atom Nanoelectronics, Pan Stanford Publishing, 2013.

[10] K. Mylvaganam, L. C. Zhang, P. Eyben, J. Mody, and W. Vandervorst, "Evolution of metastable phases in silicon during nanoindentation: mechanism analysis and experimental verification", Nanotechnology, vol. 20, 305705, pp. 8, 2009 .

[11] Athikulwongse, K., et al. *Stress-driven 3D-IC placement with TSV keep-out zone and regularity study*. in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. 2010.

[12] Ieong, M., et al., *Silicon Device Scaling to the Sub-10-nm Regime.* Science, 2004. **306**(5704): p. 2057-2060.

[13] T. Kothmayr, C. Schmitt, W. Hu, M. Brünig, and G. Carle, "DTLS based security and two-way authentication for the Internet of Things," Ad Hoc Networks, vol. 11, pp. 2710-2723, 2013.

[14] Y.S. Obeng, J. J. Kopanski, J.-J. Ahn, "Authentication article and process for making same, Patent number: US 10152666

[15] Vettiger, P., et al., *The "millipede" - nanotechnology entering data storage.* IEEE Transactions on Nanotechnology, 2002. **1**(1): p. 39-55.

Ahn, Jungjoon; Kopanski, Joseph; Obeng, Yaw S.; Kim, Jihong. "Deterministic Tagging Technology for Device Authentication." Presented at IEEE International Conference on Integrated Circuit Design and Technology (ICICDT), Dresden, DE. September 15, 2021 - September 17, 2021.

Wireless Time Sensitive Networking for Industrial Collaborative Robotic Workcells

Susruth Sudhakaran*, Karl Montgomery†, Mohamed Kashef†, Dave Cavalcanti* and Richard Candell†
* *Intel Labs, Intel Corporation*, Hillsboro, Oregon, USA, Email: {susruth.sudhakaran, dave.cavalcanti}@intel.com
† *Intelligent Systems Division, Engineering Laboratory, National Institute of Standards and Technology (NIST)*,
Gaithersburg, Maryland, USA, Email: {karl.montgomery, mohamed.kashef, richard.candell}@nist.gov

***Abstract*—In this paper, we describe a collaborative robotic workcell testbed enabled by Wireless Time Sensitive Networking (WTSN) technologies and discuss deployment, performance measurement, and management guidelines challenges. We detail the methodologies for implementing and characterizing the performance of key WTSN capabilities (time synchronization and time-aware scheduling) over IEEE 802.11/Wi-Fi. We deployed WTSN capabilities on the National Institute of Standards and Technology (NIST) collaborative robotic workcell testbed consisting of two robotic arms that emulates a material handling application, known as machine tending. We further explore configurations and measurement methodologies to characterize Quality of Experience (QoE) of this use case and correlate it to the performance of the wireless network.**

***Index Terms*—Wireless TSN, IEEE 802.11, Collaborative Robotics**

I. Introduction

Sharing computing and network resources between physical Operational Technology (OT) and digital Information Technology (IT) in smart manufacturing applications needs strictly time-synchronized and deterministic low latency communications [1]. Time Sensitive Networking (TSN) [2] and its wireless counterpart [3] are developed to achieve precise time synchronization and timeliness in a network shared by time-critical traffic and other types of traffic. This paper studies the feasibility of a wireless collaborative robotic workcell application enabled by wireless TSN.

A collaborative robotic workcell testbed was constructed at the National Institute of Standards and Technology (NIST), with service requirements that are representative of a typical machine tending application. The testbed characterizes deterministic and reliable communication needs between workcell components. The testbed baseline design over a wired network has been introduced in an earlier publication [4], while experiments with a 2x2 Multiple-Input Multiple-Output (MIMO) wireless IEEE 802.11ac technology is added in this work.

Extension activities of TSN capabilities [5] to wireless domain are described in [6]. Since Wi-Fi is an IEEE 802 Local Area Network (LAN) technology, TSN link layer capabilities can be mapped seamlessly from Ethernet to Wi-Fi. However, achieving the same wired time synchronization performance over IEEE 802.11 involves many open research questions. In this paper, we focus on characterizing and assessing the performance of two major TSN features that have been extended into IEEE 802.11, namely, time synchronization (IEEE 802.1AS) and time-aware shaping (IEEE 802.1Qbv).

II. Wireless TSN Overview

In this section, we introduce an overview of Wireless TSN (WTSN) capabilities and various system components. Fig. 1 illustrates a typical hybrid TSN network architecture where TSN capabilities in the wired segment is extended into a Wi-Fi segment of the network. It is assumed that the network is centrally managed, which is the case in most industrial internet of things (IoT) deployments that are relevant for the applications considered in this paper.

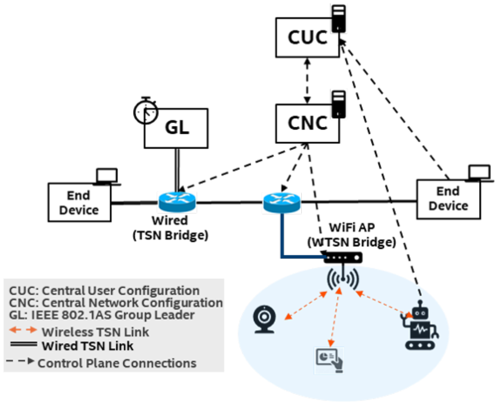

CUC
CNC
GL
End Device
Wired (TSN Bridge)
WiFi AP (WTSN Bridge)
End Device
CUC: Central User Configuration
CNC: Central Network Configuration
GL: IEEE 802.1AS Group Leader
Wireless TSN Link
Wired TSN Link
Control Plane Connections

Fig. 1. Wired-Wireless hybrid TSN network architecture.

In a TSN network, every traffic stream is centrally managed and configured. This function is performed by two functional entities namely, the Central User Configuration (CUC) and the Central Network Configuration (TSN-CNC[1]) as specified by IEEE 802.1Qcc. The CUC collects traffic stream information from all the end devices and provides the information to the TSN-CNC. The TSN-CNC, using its discovered knowledge of the network topology, configures resources on each network element on the path to meet the timing requirements of the traffic streams. This may include configuring IEEE 802.1Qbv schedules at the bridges in the infrastructure.

Achieving precise time synchronization across all the devices in the network is foundational to any TSN capable

[1] According to the TSN standard, this entity is abbreviated as CNC but since we have other entities having the same abbreviation in this paper, we will henceforth refer to this entity as TSN-CNC.

Sudhakaran, Susruth; Montgomery, Karl; Hany, Mohamed; Cavalcanti, Dave; Candell, Rick. "Wireless Time Sensitive Networking for Industrial Collaborative Robotic Workcells." Presented at 17th IEEE International Conference on Factory Communication Systems (WFCS), Linz, AT. June 09, 2021 - June 11, 2021.

network. The IEEE 802.1AS standard is the protocol defined for time distribution in a TSN network and it can operate over Ethernet and Wi-Fi (IEEE 802.11) [7]. The IEEE 802.11 specification has defined mechanisms to support the propagation of time as specified in IEEE 802.1AS. These mechanisms are based on exchange of precisely timestamped action frames between nodes to propagate a reference time. The WTSN implementation [6], used in this paper, uses the Timing Measurements (TM) feature in the IEEE 802.11 specification.

Another fundamental TSN feature is Time Aware Shaping which enables delivery of time-critical data within deterministic time windows, without being impacted by other background/interfering traffic sharing the network. Time Aware Shaping implements a time-division multiple access (TDMA) scheme of scheduling packets by giving packets of different traffic classes access to the communication medium within different time slots [2]. In this paper we evaluate our implementation of the concept of IEEE 802.1Qbv over IEEE 802.11 using an industrial workcell testbed.

III. NIST Collaborative Robotic Workcell Testbed

In this section, we present a brief description of the industrial wireless testbed design. We also present the modifications performed to the testbed to enable the WTSN. We use a collaborative workcell testbed, which was developed and implemented at NIST. We extended the testbed with WTSN and precise measurement capabilities to be able to characterize and measure the use case's physical and network performance. For brevity, the design of the testbed itself, the equipment used, as well as the information data flows are detailed in [4], and are not included in this paper. This paper will discuss the modifications made to enable WTSN.

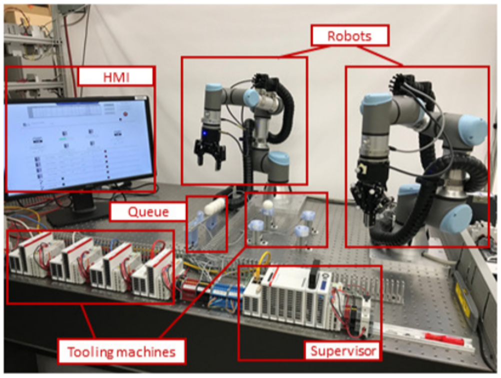


Fig. 2. Collaborative robotic workcell testbed.

At a high level, the industrial wireless testbed emulates a collaborative industrial workcell, which is shown in Fig. 2. There are two robots, operator (OPT) and inspector (INS), which move and inspect parts within the work zone. There is a supervisor programmable logic controller (PLC) that coordinates the operations in the testbed by sending commands to the robots, as well as receiving inputs from the four computerized numerical control (CNC) emulators, which are able to sense the presence of the acrylic balls in the work zone. A grand leader clock (GL) is used to synchronize all times of measurement devices and the wireless access point. A previous work that uses a similar testbed configuration and implements a graph database approach for performance evaluation without TSN can be found in [8].

To enable WTSN, wireless networking is introduced using a Wi-Fi access point (AP) and two Wi-Fi stations for each robot. These three wireless nodes are equipped with a WTSN software stack, which extends TSN features into the Wi-Fi domain. These nodes are Intel-based next unit of computing (NUC) systems (Onlogic ML100G-51) equipped with an Intel 9260 IEEE 802.11ac Wi-Fi card [9]. The AP node synchronizes its time with the GL in the network over IEEE 1588 Precision Time Protocol (PTP). The AP propagates this time over Wi-Fi to the two stations using the IEEE 802.1AS time synchronization protocol.

IV. Wireless TSN Evaluation

A. Measurement setup and methodology

To enable measurement and comparative analysis of WTSN, the nodes described in the previous section also function as measurement probes capturing and measuring wireless time synchronization data. Wired Ethernet Test Access Points (TAPs) installed at strategic points in the network, shown in Fig. 3, capture all network packets sent between the nodes at multiple points. Multiple 4-port Ethernet PCIe cards simultaneously capture traffic from the TAPs, shown in Fig. 3. Data is also captured from the robots through its real-time-data-exchange (RTDE) interface. The measurement methodology involves running the machine tending use case in the presence of various levels of interfering traffic and evaluating network and application performance with and without WTSN. The complete measurement setup is illustrated in Fig. 3.

To measure the network performance metrics, difference in the capture time of the same traffic at multiple points is leveraged. Moreover, the time synchronization errors between any node and the GL range from less than 1 microsecond, in the case of wired nodes, to less than 100 microseconds in the case of wireless nodes with a 99% confidence interval.

B. Key Performance Indicators (KPIs)

The KPIs and corresponding metrics were chosen while keeping in mind the goal of evaluating the performance of WTSN on the testbed. In order to evaluate the overall performance, different types of KPIs are needed. The metrics under the network performance KPIs will provide a measure of how well the network is performing with respect to the time sensitive requirements of the application. Similarly, the metrics under the application KPIs provide measures of the efficiency of the application. The metrics measured in this work, and their description are summarized in Table. I.

Sudhakaran, Susruth; Montgomery, Karl; Hany, Mohamed; Cavalcanti, Dave; Candell, Rick. "Wireless Time Sensitive Networking for Industrial Collaborative Robotic Workcells." Presented at 17th IEEE International Conference on Factory Communication Systems (WFCS), Linz, AT. June 09, 2021 - June 11, 2021.

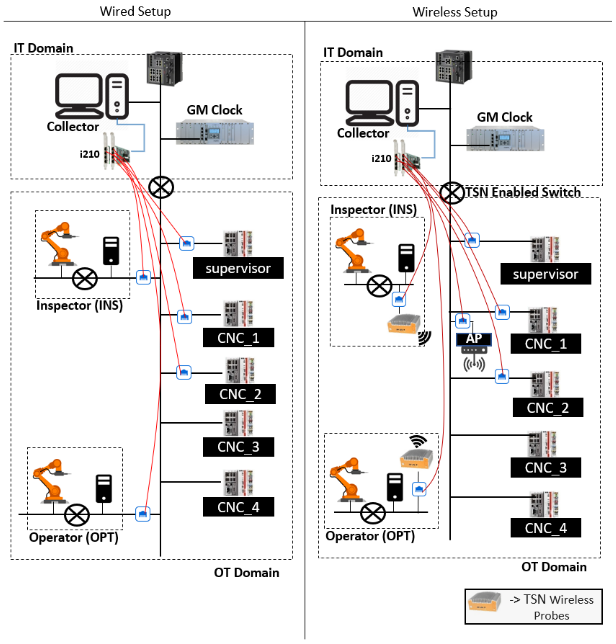


Fig. 3. Wireless TSN measurement setup of workcell testbed.

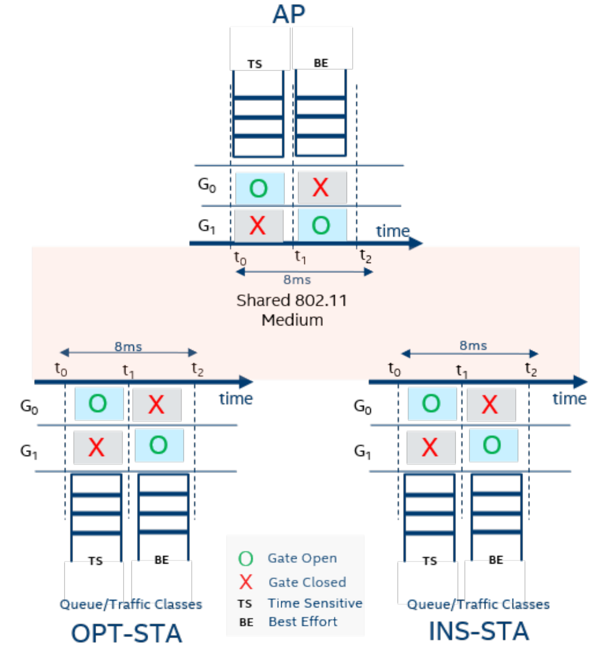


Fig. 4. WTSN Schedule in the wireless medium.

TABLE I
KPI METRICS

KPI	Description
Packet Delivery Ratio (PDR)	PDR is defined as the ratio of data packets delivered within a defined latency threshold between key points in the network.
Latency Cumulative Distribution Function (CDF)	The Latency CDF captures the distribution of latency across all data packets observed for a network flow.
Idle Time	The amount of time the robots have spent idling due to delays in updating state changes.

C. Time-aware schedule for wireless domain

The traffic between the PLC and the two robots are the two Time Sensitive (TS) traffic streams in the use case. In order to protect this traffic from interference due to any best effort (BE) traffic, the shared 802.11 link implements a time-aware (802.1Qbv) based schedule with protective slots for the TS traffic. This schedule is illustrated in Fig. 4. The schedule repeats every 8 ms to match the frequency of the two TS streams and each 8ms period is further divided into slots for TS traffic (5ms) and BE (3ms) traffic. This schedule is synchronized across all the nodes in the shared link in order to create protective periods for the TS traffic. The schedule also takes advantage of the inherent staggering between the two TS traffic streams.

The start of the application's use case and the start of the schedules are synchronized as close as possible through automated synchronization scripts. In an ideal TSN network this schedule and synchronization will be coordinated centrally by the CUC and TSN-CNC, as discussed previously. Although the protocols for coordination and resource management have been standardized, implementation and optimization of these entities are still an active area of research, especially in the wireless domain. It should be noted that the capacity available to BE traffic is reduced as a result of enabling the schedule as the BE traffic gets only 37.5% of the cycle period for transmitting. On top of that, some of that bandwidth will be further consumed by opportunistic shaping, or guard band, hence the effective rate of the medium is reduced. Note that this is an acceptable trade-off in TSN networks as the primary goal is to guarantee determinism for time-critical traffic.

V. Results

In this section, we will discuss the results produced from the testbed. The measurement methodology is highlighted in section IV-A. Fig. 5 shows the PDR for packets with a measured latency of less than 5 ms. We can see a significant drop in PDR without TSN enabled, which is related to the amount of interference traffic. We can see that enabling the TSN schedule is able to bring the overall latency profile of the time sensitive streams closer to the wireless baseline benchmark (99.8%).

In Fig. 6 we show the cumulative density function (CDF) comparison of the latency distribution of operational traffic in the presence of worst-case interference traffic (20 Mbps) when the TSN schedule is both enabled and disabled. When the TSN schedule is enabled, more than 99% of time sensitive packets experience a bounded latency of less than 5ms. When

Sudhakaran, Susruth; Montgomery, Karl; Hany, Mohamed; Cavalcanti, Dave; Candell, Rick. "Wireless Time Sensitive Networking for Industrial Collaborative Robotic Workcells." Presented at 17th IEEE International Conference on Factory Communication Systems (WFCS), Linz, AT. June 09, 2021 - June 11, 2021.

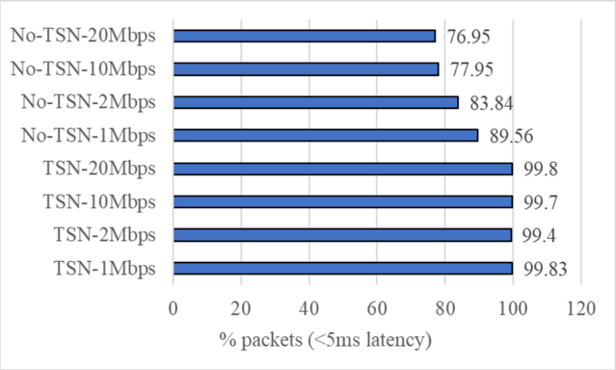


Fig. 5. Packet Delivery Ratio for latency < 5ms with TSN and no-TSN cases with varying levels of interference traffic.

the TSN schedule is disabled, the percentage of packets within the latency bound of 5ms decreases to approximately 77%.

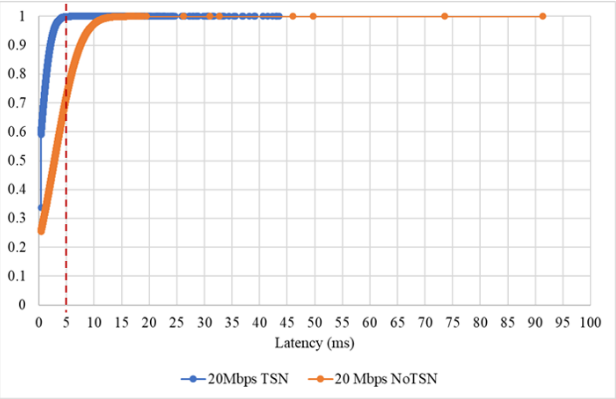


Fig. 6. Latency CDF of packets with and without TSN enabled.

The increase in the percentage of packets experiencing latency outside of tolerable bounds is also reflected in the application's performance, as shown in Fig. 7.

With TSN, the percentage of time in idle experienced by the operator robot is lower compared to the case when TSN is disabled. When TSN is disabled, there is competing BE traffic in the network with no protection for the time sensitive traffic, which increases latency, and consequently, the idle time of the operator robot as is takes more time to receive commands from the supervisor PLC. The decrease in the wireless link delay from TSN being enabled increases the efficiency of the use case, which can be desired to increase the production rate in an industrial setting where collaborative robots accomplish supervisory tasks over wireless.

VI. Conclusions

In this paper, wireless TSN capabilities were used to deliver low latency communications for an industrial collaborative robotics use case. A detailed analysis of the latency and its correlation to the overall use case efficiency was presented. There are many challenges that should be addressed in order to fully support TSN in the wireless domain. We hope to address these challenges and revisit these experiments with new scheduling and TSN capabilities in Wi-Fi 6 and beyond.

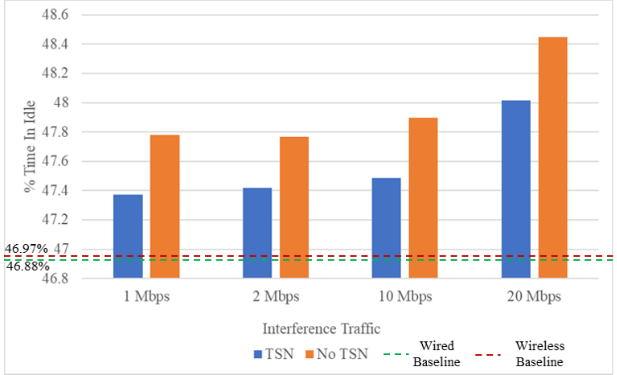


Fig. 7. Operator (OPT) Idle time across all scenarios.

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] Y. Liu, M. Kashef, K. B. Lee, L. Benmohamed and R. Candell, "Wireless Network Design for Emerging IIoT Applications: Reference Framework and Use Cases," in Proceedings of the IEEE, vol. 107, no. 6, pp. 1166-1192, June 2019.

[2] IEEE 802.1 Time Sensitive Networking (TSN) Task Group: https://1.ieee802.org/tsn/.

[3] M. Eisen, M. M. Rashid, A. Ribeiro, and D. Cavalcanti. "Scheduling Low Latency Traffic for Wireless Control Systems in 5G Networks." arXiv preprint arXiv:1910.13587 (2019).

[4] Y. Liu, R. Candell, M. Kashef, and K. Montgomery, "A collaborative workcell testbed for industrial wireless communications—the baseline design," 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), pp 1315–1321, June 2019.

[5] J. Farkas, L. L. Bello and C. Gunther, "Time-Sensitive Networking Standards," in IEEE Communications Standards Magazine, vol. 2, no. 2, pp. 20-21, JUNE 2018.

[6] D. Cavalcanti, J. Perez-Ramirez, M. M. Rashid, J. Fang, M. Galeev and K. B. Stanton, "Extending Accurate Time Distribution and Timeliness Capabilities Over the Air to Enable Future Wireless Industrial Automation Systems," in Proceedings of the IEEE, vol. 107, no. 6, pp. 1132-1152, June 2019.

[7] IEEE 802.1AS Specification. https://standards.ieee.org/standard/802_1AS-2011.html

[8] M. Kashef, Y. Liu, K. Montgomery, and R. Candell, "Wireless Cyber-Physical System Performance Evaluation Through a Graph Database Approach." ASME. J. Comput. Inf. Sci. Eng., vol. 21, No. 2, April 2021, Published Online: October 16, 2020.

[9] Intel Wi-Fi 5 Discrete Client Reference, https://ark.intel.com/content/www/us/en/ark/products/99445/intel-wireless-ac-9260.html, Accessed: 2/11/2021.

Sudhakaran, Susruth; Montgomery, Karl; Hany, Mohamed; Cavalcanti, Dave; Candell, Rick. "Wireless Time Sensitive Networking for Industrial Collaborative Robotic Workcells." Presented at 17th IEEE International Conference on Factory Communication Systems (WFCS), Linz, AT. June 09, 2021 - June 11, 2021.

Feature Extraction and Classification for Communication Channels in Wireless Mechatronic Systems

Jing Geng*, Mohamed Kashef†, Richard Candell†, Yongkang Liu†, Karl Montgomery†, Shuvra S. Bhattacharyya*
* Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
† Intelligent Systems Division, Engineering Laboratory, National Institute of Standards and Technology, USA
Email: jgeng@umd.edu, {mohamed.kashef, richard.candell, yongkang.liu, karl.montgomery, }@nist.gov, ssb@umd.edu

Abstract— **For accurate characterization and evaluation of wireless mechatronic systems, effective modeling of wireless communication channels is of paramount importance, especially to simulation-oriented methods. Conventional simulation methods employ mathematical models to abstract details of prototype channels. Although such mathematical models often have rigorous theoretical underpinnings, they can be weak in capturing complex environmental characteristics and complex forms of diversity that are exhibited in industrial communication environments. To address this problem, we develop, in this paper, a new approach to deriving effective simulation models for industrial communication channels. Our approach involves field measurements from actual wireless mechatronic environments together with feature extraction from the measurements, and data-driven classification based on the extracted features. Our approach leads to a general framework for simulating wireless mechatronic systems in a way that realistically incorporates the complex channel characteristics of these systems.**

I. Introduction

In recent years, integrating wireless communications capabilities into industrial mechatronic systems has attracted great interest due to promising advantages offered by wireless communications [1]. This integration brings about highly complex design spaces, which we refer to as *wireless-integrated factory system* (*WIFS*) design spaces, involving interactions among physical factory layout, control algorithms, and wireless communication networks (e.g., see [2]). General and effective simulation methods are needed for WIFS environments and other types of complex mechatronic environments for performance assessment of existing system designs, development of new designs, and planning of technology updates.

Many software tools have been developed to provide sophisticated capabilities for communication network simulation. However, traditional radio frequency (RF) wireless channel models that are used with such tools have significant limitations in the context of simulating WIFS environments. Such models typically utilize mathematical formulations associated with collected data or power delay profiles (PDPs) obtained from third-party studies (see Section II). WIFS simulations with such models may not accurately reflect actual system performance since the models do not take into account harsh conditions and other distinguishing characteristics of RF channels in industrial communication environments. Such characteristics arise, for example, from vibration of machinery, and the presence of metal objects (e.g., see [3]), which can have a major impact on communication performance.

In this paper, we build upon our recent work that introduced a new link layer simulation approach for integrating field measurements into WIFS simulation [4]. We refer to that method as the Measurement-based Channel Library Generator (MCLG) method, which involves constructing channel library modules, in the form of PER/SNR (packet error rate / signal-to-noise ratio) tables, from field measurements in a systematic manner. However, the MCLG method has a couple of limitations. First, the clustering algorithm used based on [5] is computationally intensive, and the running time can become very large with large measurement datasets. Second, the derived channel libraries correspond specifically to the measured environment for the specific measured communication paths. In large source environments, it is unrealistic to measure channel impulse responses (CIRs) for all possible communication paths.

In this paper, we develop significant improvements to the MCLG method that address the two limitations described above, and allow the method to be used more generally, in a broader class of WIFS design space exploration scenarios for a given source environment. We refer to the improved version of the MCLG method as the Generalized Measurement-based channel Library Generator (GMLG) method, and we refer to our prototype implementation of the method as GMLG.

II. Related Work

Candell et al. discuss challenges involved in employing industrial wireless technology in mechatronic systems, and present guidelines for addressing those challenges [6]. A large body of literature addresses the modeling of wireless communication channels with emphasis on different communication modeling concerns (e.g., see [7], [8]).

Various works have investigated the modeling of communication channels in more challenging communication environments. For example, Abbas et al. performed comparisons between simulations and measurements to evaluate vehicle-to-vehicle communication channel parameters [9]. Peil et

Geng, Jing; Hany, Mohamed; Candell, Rick; Liu, Yongkang; Montgomery, Karl; Bhattacharyya, Shuvra. "Feature Extraction and Classification for Communication Channels in Wireless Mechatronic Systems." Presented at 17th IEEE International Conference on Factory Communication Systems (WFCS). June 09, 2021 - June 11, 2021.

al. developed channel models using wireless propagation characteristics in an industrial environment [10]. Other works emphasize application of existing simulation frameworks to design challenges in specific application areas (e.g., see [2], [11]–[13]).

The GMLG method presented in this paper differs from previous works, including those summarized above, in its emphasis on systematically incorporating field measurements into simulation of wireless mechatronic systems, and its application of feature clustering to generalize the types of communication channels that can be simulated using a given set of field measurements.

III. Methods

Fig. 1 provides a block diagram illustration of the GMLG method. Input to GMLG consists of a set $S_c = \{C_1, C_2, \ldots, C_m\}$ of CIR measurements from distinct physical communication paths. The output includes a set $S_k = \{K_1, K_2, \ldots, K_n\}$ of representative channels along with a PER/SNR table $T(K_i)$ that characterizes the communication conditions represented by each K_i. The output also includes a machine learning model M that maps features extracted from communication paths in the source environment into representative channels. The primary components of MCLG that are reused in GMLG are the Link Level Simulator and Pre-processing block (see Fig. 1). A new clustering process is devised in GMLG to be more computationally efficient and to support the objective of deriving the classification model M.

In summary, the GMLG method processes CIR measurements to produce as output a triple (S_k, T, M), where T provides a mapping of the representatives in S_k into corresponding PER/SNR tables, and M provides a mapping of features extracted from arbitrary communication paths into S_k. Each ordered pair $(K_i, T(K_i))$ is referred to as a *library module* that is generated by the $GMLG$ method, and can be plugged into system-level simulators to aid in exploring WIFS design spaces.

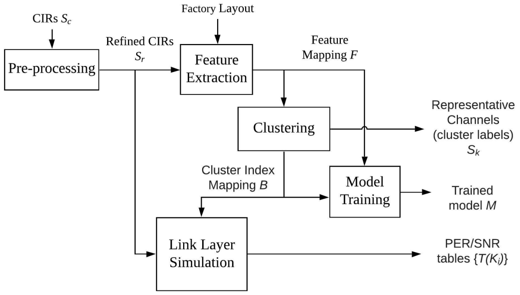


Fig. 1. An illustration of the GMLG method.

A. Feature Selection and Extraction

The input to the Feature Extraction block in Fig. 1 is a set of *refined* CIRs $S_r = \{R_1, R_2, \ldots, R_n\}$ that is produced by the Pre-processing block. Operations performed in the Pre-processing block include noise filtering, intra-CIR compression, and peak power determination, which are discussed in [4]. The output of the Feature Extraction block is a feature mapping $F : S_r \rightarrow S_f$, where for each $R_i \in S_r$, $F(R_i)$ is the feature vector extracted from the refined CIR R_i.

At this point, it is useful to distinguish between two types of features that are employed in GMLG — *external* and *system* features. External features are those that are input to the generated machine learning model M when applying M to higher level simulations by users. It should be possible to derive external features conveniently from data associated with the layout of the source environment. That way, new communication paths can be predicted by the model M even though they are not included in the measurements. On the other hand, system features are the features that are used to train the model M. In general, the system feature set includes all of the external features along with zero or more additional features. System features can be defined in terms of field measurement data — for example, they can be extracted from the CIRs collected to form the feature vector S_f. In GMLG, we have incorporated three external features, which are the path distance, line of sight (LOS) / non-LOS (NLOS) indicator, and Rician K Factor, with two additional features — mean delay and root mean square (RMS) delay spread — that are included in the system feature set.

B. Clustering

The Clustering block in Fig. 1 partitions refined CIRs into groups of related CIRs. Each group or *cluster* corresponds to a distinct channel library module that is generated by GMLG. A distinguishing aspect of GMLG is that clustering is performed in the feature space — that is, the clustering algorithm operates on the feature vectors rather than on the refined CIRs themselves. This enables much more efficient clustering since the feature vectors are highly compressed representations of the refined CIRs.

A wide variety of clustering techniques can be used in the GMLG method. Different clustering algorithms may be chosen depending, for example, on the spread of data points for each feature. In GMLG, we apply spectral clustering [14], which is known for its effectiveness in handling categorical data, and for its ability to handle inseparable data and derive non-convex clusters.

C. Clustering Assessment and Weight Optimization

For assessment of clustering solutions, we consider two different metrics, the silhouette coefficient (e.g., see [15]), and the average feature variance. In our context, a cluster corresponds to a set of refined CIRs that are grouped together by the Clustering block (Fig. 1), and a clustering solution corresponds to a partitioning of all refined CIRs into disjoint subsets (candidate clusters).

To assess the quality of clustering solutions in GMLG, we employ a composite metric that is formed of different sub-metrics, which are listed below. For a given sub-metric, we refer to the feature set used in assessment as the *evaluation feature set*, which is a subset of the five system features.

- Γ_1 is defined as the silhouette coefficient using an evaluation feature set that consists only of the K factor, LOS/NLOS indicator, and path distance (external feature set of GMLG).
- Γ_2 is defined as the silhouette coefficient using an evaluation feature set that consists only of the K factor, mean delay, and RMS delay spread. We refer to these features as ***quality assessment features*** since they are representative features for capturing channel characteristics from a measured CIR. The sub-metric Γ_2 is the main sub-metric for guiding the tuning process of clustering in GMLG.
- $\Gamma_3, \Gamma_4, \Gamma_5$ are variance metrics for the quality assessment features — K factor, mean delay, and RMS delay — respectively.

Although the external feature set is used for feature clustering, all five metrics $\Gamma_1, \Gamma_2, \ldots, \Gamma_5$ are used to guide the tuning process for feature weights when invoking spectral clustering. In other words, GMLG executes spectral clustering iteratively using different relative weightings of the features. Intuitively, the output solution B is selected as a solution that maximizes Γ_2, while retaining reasonable performance on the other four metrics.

D. Model Training

The clustering solution provided by the Clustering block is used to train a machine learning model, as illustrated by the Model Training Block in Fig. 1. The objective is to derive a trained machine learning model M that can map novel paths, based on estimated features from those paths, into representative channels.

The input to Model Training are the feature vectors $F(R_1), F(R_2), \ldots, F(R_N)$ for the refined CIRs along with the corresponding cluster indices $B(R_1), B(R_2), \ldots, B(R_N)$. The cluster indices are used as labels for supervised learning based on feature vectors. Intuitively, the model is trained to predict a well matched representative channel from a given feature vector composed of the external features.

There are many types of classifiers that can be applied within the GMLG method for the derivation of M. In GMLG, we apply the k-nearest-neighbors (KNN) algorithm [16], which is capable of handling classification functions involving categorical output results.

IV. Experiments

The set of field measurements that we use in our experiments consists of CIRs obtained from a measurement campaign from an automotive factory site performed by National Institute of Standards and Technology (NIST) [17]. We implement clustering and model training in Python (Version 3.7.4) and make use of the `scikit-learn` package [18] (Version 0.22).

A. Feature Clustering Results

The automotive factory site dataset includes 41,700 CIRs, which are later reduced to 1,524 refined CIRs after preprocessing. The results of the clustering process in GMLG on the automotive factory dataset are illustrated in Fig. 2. The figure provides a perspective on how clusters are formed in relation to different pairs of features. Each data point in each of the three plots corresponds to a feature vector that is projected onto the two-dimensional subspace corresponding to each plot. The data points are colored differently based on which cluster they are assigned to. The results illustrate that the clustering process is effective in separating the feature vectors into distinct regions of the feature space — this is perhaps most strongly demonstrated by the plot involving the K factor and path distance.

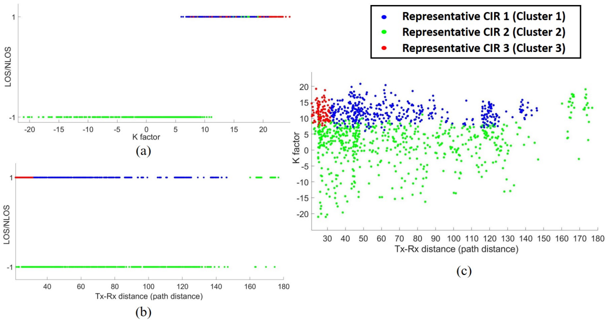


Fig. 2. Separation of feature vectors into clusters for the automotive factory dataset: (a) K factor vs. LOS/NLOS indicator, (b) path distance vs. LOS/NLOS indicator, (c) path distance vs. K factor.

B. Comparison Between AP-DTW and Feature Clustering

In this section, we present results that demonstrate significant improvements provided by the feature-based clustering method in GMLG compared to the clustering approach in MCLG, which is based on affinity propagation (AP) and dynamic time warping (DTW).

Because the computation time for the AP+DTW approach grows very rapidly with the dataset size, we perform the comparison experiment in this section using a smaller dataset with 254 processed-CIRs. This dataset is derived by using a larger downsampling factor during pre-processing.

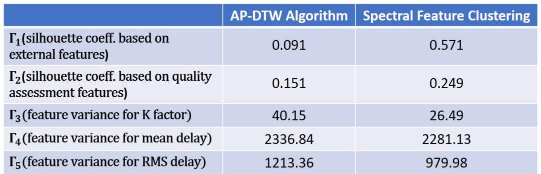
The improvement in computational speed is significant. We found that the APT+DTW method required an average of 4,217 seconds with standard deviation $\sigma = 194.0$, while GMLG required an average of 23.47 seconds with $\sigma = 5.64$. These results represent a speedup of 179.7X. The execution time measurements were averaged over 20 runs.

Table I compares the quality of the derived clustering solutions between AP+DTW and GMLG. The results show significant improvements in terms of all of the five evaluation metrics $\{\Gamma_i\}$ that were defined in Section III-C, especially for the silhouette metric Γ_1 and the variance metric Γ_3.

C. Classifying Novel Paths

As motivated in Section I, an important capability of the GMLG method is the ability to classify new communication paths (i.e., paths that do not correspond to any of the paths covered in the field measurements that are used to construct

TABLE I
COMPARISON OF CLUSTERING SOLUTIONS.



	AP-DTW Algorithm	Spectral Feature Clustering
Γ_1 (silhouette coeff. based on external features)	0.091	0.571
Γ_2 (silhouette coeff. based on quality assessment features)	0.151	0.249
Γ_3 (feature variance for K factor)	40.15	26.49
Γ_4 (feature variance for mean delay)	2336.84	2281.13
Γ_5 (feature variance for RMS delay)	1213.36	979.98

the clusters). We evaluate this capability in our experiments by applying the derived classification model M on a *testing dataset* that is extracted from the original automotive factory measurement dataset, and that is excluded from the set of CIRs that is used in clustering and model training in GMLG. The paths in the testing dataset can be viewed as novel communication paths.

In this experiment, the classification labels produced in the experiment presented in Section IV-A are used as the ground truth. We then evaluate how accurately the model M, which is produced by GMLG, predicts the cluster label from the highly compressed (feature-based) representation that M operates on. We use 80% of the dataset as input to GMLG, which is configured to generate three representative channels. The remaining 20% of the dataset is used for testing.

The results of this experiment show that the number of misclassifications is only 3 out of 305 total testing instances, for a testing accuracy of 99.0%. The experiment therefore demonstrates the potential of the GMLG method to produce high accuracy mappings of novel communication paths into representative channel library modules

V. CONCLUSIONS

This paper has introduced a new approach for integrating field measurements into the modeling and simulation of mechatronic systems that are integrated with wireless communication capability. Novel aspects of the approach include feature extraction and feature clustering, which allow for derivation of representative channels and associated channel library modules in an efficient manner from highly compressed, feature-based representations. Moreover, the derived clusters are used to train a classification model, which can be used to classify novel communication paths (not represented in the field measurements) into the most representative channel library modules for simulation. In the experiments presented in the paper, key parameters of the proposed approach, such as the downsampling factor applied to CIRs, and the number of clusters to generate, were derived empirically. A useful direction for future work is the development of automated methods and supporting tools for setting these parameters.

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] 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*, 2019, pp. 1–8.

[3] 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.

[4] J. Geng, H. Li, M. Kashef, Y. Liu, R. Candell, and S. S. Bhattacharyya, "Integrating field measurements into a model-based simulator for industrial communication networks," in *Proceedings of the IEEE World Conference on Factory Communication Systems*, 2020, pp. 1–8.

[5] 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.

[6] R. Candell, M. Kashef, Y. Liu, K. B. Lee, and S. Foufou, "Industrial wireless systems guidelines: Practical considerations and deployment life cycle," *IEEE Industrial Electronics Magazine*, vol. 12, no. 4, pp. 6–17, 2018.

[7] J. Medbo and P. Schramm, "Channel models for HIPERLAN/2 in different indoor scenarios," Ericsson Radio Systems AB, Tech. Rep. 3ERI085B, 1998.

[8] V. Erceg, "TGn channel models," IEEE P802.11 Wireless LANs, Tech. Rep. IEEE 802.11-03/940r4, 2004.

[9] T. Abbas *et al.*, "Simulation and measurement-based vehicle-to-vehicle channel characterization: Accuracy and constraint analysis," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 7, pp. 3208–3218, 2015.

[10] J. Peil *et al.*, "Channel modeling and performance of Zigbee radios in an industrial environment," National Institute of Standards and Technology, Tech. Rep., September 2016.

[11] M. A. Ahmed, W.-H. Yang, and Y.-C. Kim, "Simulation study of communication network for wind power farm," in *Proceedings of the International Conference on Information and Communication Technology Convergence*, 2011.

[12] 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.

[13] 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.

[14] M. C. V. Nascimento and A. C. P. L. F. de Carvalho, "Spectral methods for graph clustering — a survey," *European Journal of Operational Research*, vol. 211, no. 2, pp. 221–231, 2011.

[15] P. J.Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[16] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[17] "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.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A Framework for the Composition of IoT and CPS Capabilities

Khalid Halba[1,2,3], Edward Griffor[2], Ahmed Lbath[1] and Anton Dahbura[3]

[1]Grenoble Alpes University, Grenoble, France
{khalid.halba , ahmed.lbath}@univ-grenoble-alpes.fr
[2]National Institute of Standards and Technology, Gaithersburg, Maryland, USA
{khalid.halba , edward.griffor}@nist.gov
[3]Johns Hopkins University, Baltimore, Maryland, USA
{khalba1 , adahbura1}@jhu.edu

***Abstract*—By 2030, over a half trillion devices will be connected to the internet. With so many devices providing a wide range of features, there is a need for a framework for innovation and reuse of Internet of Things (IoT) and Cyber-Physical Systems (CPS) capabilities. Such framework should facilitate the composition of capabilities and provide stakeholders means to reliably model and verify compositions. An IoT and CPS Composition Framework (ICCF) is proposed to achieve this goal. ICCF is based on the NIST CPS framework composition guidelines, intuitive composition semantics inspired from the mPlane protocol, and strong formal verification capabilities of the Temporal Logic of Actions (TLA) formal descriptors and tools. This paper demonstrates why such framework, semantics, and formal specification and verification components form a powerful and intuitive composition framework that satisfies different stakeholders concerns. To achieve this purpose, semantics and formal specification of the composition algebra were provided, a well-being composite capability within a smart building was specified, its prototype model in a formal verification tool was run, an analysis of the results of symbolic execution quantitatively and qualitatively was performed, and assessment of the trustworthiness of the composition was done. Lastly, implementation details were provided and proposed extensions to other domains such as smart transportation and smart health were discussed.**

***Index Terms*—Framework, IoT, CPS, ICCF, Capability, Algebra, Composition, Trustworthiness.**

I. Introduction and paper plan

IoT or CPS capabilities composition is the process of generating a value-added capability based on atomic measurements or services. Throughout this paper, IoT and CPS can be used interchangeably [1]. A framework for capabilities composition that addresses different stakeholders concerns would serve as a foundation for open innovation and re-purposing of IoT and CPS capabilities of an expected half-trillion IoT and CPS devices by 2030 [2]. Examples of target domains include smart buildings (well-being), transportation (safety of autonomous vehicles), and healthcare (autonomous ventilators output). Verifying such novel compositions and making sure their deployment won't cause errors is crucial for a trustworthy implementation. For this reason, there is a need for a framework for composing IoT or CPS systems capabilities regardless of their complexity. This work proposes an **I**oT and **C**PS **C**omposition **F**ramework (**ICCF**) that addresses these goals. Such a framework should lay the groundwork for composition and trustworthiness assessments, use straightforward semantics to help developers prototype novel capabilities, and describe tools for the formal verification of such novel composite capabilities. Capabilities composition taxonomy involves formal, technical, and Quality of Service (QoS) components. Together these components provide foundations for representing, processing, and generating sought-after capabilities by diverse stakeholders while preserving properties of interest to those stakeholders. The focus of this paper is on formal components of the proposed framework, technical and QoS aspects remain for future work. Formal components of ICCF include i) foundations for composition ii) semantics for capabilities, interactions, and compositions iii) formal specification languages and formal verification tools used to translate the semantics of composition into specifications for performing model checking and trustworthiness assessment via assertions or deadlock analysis. The contributions of this paper are organized as follows: in section II, related work on previous efforts is provided. In section III: i) criteria based on which the selection of the building blocks of the proposed composition framework is defined, ii) existing frameworks, semantics, and formal verification techniques are compared to select formal components that best

Halba, Khalid; Griffor, Edward; Dahbura, Anton; Lbath, Ahmed. "A Framework for the Composition of IoT and CPS Capabilities." Presented at COMPSAC 2021: Intelligent and Resilient Computing for a Collaborative World 45th Anniversary Conference. The 4th IEEE International Workshop on Smart IoT Sensors & Social Systems for eHealth & Well-Being Applications (SIS-SS 2021). July 12, 2021 - July 16, 2021.

satisfy ICCF requirements. Section IV provides ICCF semantics used to describe interactions and capabilities compositions. In section V an example of composition in the IoT space is presented, its composite function is formally specified and its state space is studied following its model checking. This is followed by quantitative and qualitative analysis and an assessment of the trustworthiness of the results obtained, as well as an outlook on implementation efforts of the studied example and potential extensions to other domains. In the conclusion, a summary of contributions is presented as well as an outlook to future work.

II. Related work

Examples of IoT and CPS frameworks, environments, or standards for capabilities composition include OneM2M environment [3]. It leverages the NIST CPS framework [4] [5] [6], a comprehensive framework that provides capabilities composition guidelines including time synchronization between atomic capabilities. The OneM2M environment can use the M2M semantics provided by the industry segment that uses corresponding data. This makes its semantics domain-specific [7], a higher abstraction layer might be needed to simplify rapid prototyping of composition for different IoT and CPS domains. Fiware [8] was coupled with the IoT-A framework which supports IoT capabilities composition and semantic specification using Business Process Model and Notation (BPMN) 2.0, it also supports powerful features such as synchronous and asynchronous compositions [9]. The BMPN semantics, however, make it challenging to formally verify compositions as that involves converting the BPMN notation to the Generic Property Specification Language (GPSL formal) specification language which can improve expressiveness but might add complexity, impact performance, or limit expressiveness when converting BMPN to a graphically verifiable model such as Property Sequence Chart (PSC) [10]. For the CIM (Context Information Management) environment [11], the foundations for composition are provided by the CIM NGSI evolution framework, it uses RDF (Resource Description Framework) to semantically describe the capabilities of a system. RDF is a graph-based descriptive language, it can be converted to a formally verifiable specification such as ShEx 2.0 (Shape expression schemas v2.0) [12]. ShEx expressions can be used both to describe RDF and to automatically check the conformance of RDF data. However, ShEx checks whether RDF data respects the schema requirements as it is data-oriented not composition function-oriented. This can make it challenging to model check the system features. VITAL is another project that supports IoT-A framework, W3C SSN semantics, but recommendation on formal specification and verification languages and tools to use are not the focus of the framework [13]. The same case for FogFlow [14], an environment that leverages NGSI framework for IoT capabilities composition foundations and YAML as a capability descriptor. AWS [15] is a commercial environment for cloud services, it leveraged PlusCal semantics [16] and TLA [17] formal specification techniques to verify the correctness of properties such as fault tolerance in their storage services, but this didn’t extend to cover the microservices and IoT composition solutions such as GreenGrass [18]. To address the limitations of these frameworks and to provide a strong framework for the composition of IoT and CPS capabilities, an IoT and CPS Composition Framework (ICCF) is proposed. This framework leverages the NIST CPS framework composition and trustworthiness recommendations, uses strong semantics inspired from the mPlane protocol [19], and relies on the intuitive PlusCAL/TLA/TLA+ package to prototype, formally specify, and model check capabilities and assess their trustworthiness.

III. Formal criteria and ICCF foundations

A. Formal criteria

Satisfying the formal criteria of the ICCF framework means relying on a framework that provides capabilities composition foundations. This also requires the leverage of lightweight and expressive semantics to describe compositions and being able to translate their semantics easily to a formal specification language for building complex functions and verifying their trustworthiness. Below is a comparison of frameworks, semantics, and formal verification techniques that aims to select components providing the formal components necessary to satisfy the requirements of ICCF.

B. Comparing formal components

1) Frameworks: A framework enables composition when it takes into consideration i) concerns related to the ability of the IoT/CPS to achieve an intended purpose in the face of changing external conditions such as the need to upgrade or otherwise reconfigure an IoT/CPS to meet new conditions, needs, or objectives (adaptability), ii) concerns related to our understanding of the behavior of IoT/CPS due to the richness and heterogeneity of interactions among its components, such as the existence of legacy components and the variety of interfaces (complexity), iii) concerns related to the ability to combine IoT/CPS modular components (hardware, software, and data) to satisfy user requirements (constructivity), vi) concerns related to the ease and reliability with which an IoT/CPS component can be observed and understood (for purposes of leveraging the component's functionality) by an entity (human, machines), and v) concerns related to the ease and reliability with which an IoT/CPS component's functions can be ascertained (for purposes of leveraging that functionality) by an entity (discoverability). IoT-A is a reference model and architecture (RMA) designed to allow the generation of different IoT architectures tailored to specific scenarios. Using IoT-A with Fiware in [8] enabled the creation of architectures with different functional groups each serving a specific purpose and enabling interoperability, the composition of functions is intrinsic to Fiware using the service organization FG (functionality group) but stakeholders concerns when composing IoT capabilities aren't explicitly addressed. In [20], an OWL-based ontological framework for the opportunistic composition of IoT systems was introduced. The framework leverages holons, which are programming entities used to model distributed systems. Designing holons uses CoAMOS and A3ME ontologies. The resulting ontologies can then be converted to UML or domain-specific languages for further exploitation or composition in the IoT domain. While the capabilities discoverability or composition complexity are addressed in this framework, the adaptability of the composition isn't addressed. In [21], ISCO, (Internet of Smart City Objects), a distributed framework for service discovery and composition was introduced with three major enablers: a semantic functional description of city objects, representing physical devices or abstract services, a distributed service directory that embodies available city services for service lookup and discovery, and planning tools for selecting and chaining basic services to compose new complex services, this effort provides rich implementation aspects in the smart city context but the trustworthiness of composed capabilities isn't addressed. The NIST CPS framework [4] defines criteria that contribute to CPS composition trustworthiness taking into consideration functional, human, trustworthiness, timing, data, and composition concerns. The composition concern addresses adaptability, complexity, constructivity, and discoverability of CPS capabilities, hence, the NIST CPS framework composition foundations are leveraged to guide stakeholders concerns for composing capabilities.

2) Semantics: Semantics for capabilities composition suggest that the semantics are lightweight and expressive enough to represent capabilities, interactions, compositions, and workflows necessary to compose value-added features in IoT and CPS. The W3C Web Ontology Language (OWL) [22] is a semantic web language designed to represent complex and rich IoT capabilities, groups of things, and relations between things. However, it is more geared toward web services and it is not a lightweight approach to composing services. In [23], CyPhyML, a CPS capability description language was discussed with formal specification capabilities supported. However, the language was more geared towards a formal description of CPS systems for model checking purposes and not IoT services. mPlane semantics [19] allow the representation of value-added capabilities using a set of operations designed to facilitate service composition. These compositions can be applied to measurement environments, IoT services, and CPS. mPlane semantics are simple, expressive, and lightweight compared to other description languages investigated, as a result, it satisfies the human aspect of the NIST CPS framework.

3) Formal specification and verification: This aspect is related to the semantics discussed earlier. Building correct compositions and verifying their properties should not be a daunting experience for engineers and developers. These stakeholders should be able to easily use semantics and service descriptors to formally specify and prototype composite services. In [24] authors introduced linear logic LL based on pi-Calculus to describe and formally verify non-functional attributes such as the credibility/trustworthiness of the service composition. Linear temporal logic (LTL) was introduced in [25]. Real-Time Maude formal verification tool that is based on LTL was used to formally verify properties of interest. In [26], Directed acyclic graphs were used to formally model dynamic service discovery, invocation, and composition in opportunistic networks. Petri Nets [27] were used as an algebra to formally model services and processes where the main goal was to formally verify compliance of compositions with the ever-changing regulations on IoT. Temporal Logic of Actions formal specification was used to formally specify and verify critical properties on services within the AWS echo-system [15]. The common aspect between these formal specification languages is how difficult it is to move from description semantics to formal specification of composite services. Except for TLA, which has strong software support using PlusCAL, a high-level language that is comparable to pseudocode and which enables the fast translation of mPlane composition semantics to TLA formal specification. TLA+ is the formal tool that uses notation which is very similar to natural mathematic operations. CoQ [28] or Isabelle [29] use relatively daunting notations that are challenging for stakeholders which might impact the developer's ability to write verifiable compositions and as a result might limit innovation. Based on this comparative study of frameworks, service description semantics, and formal specification and verification

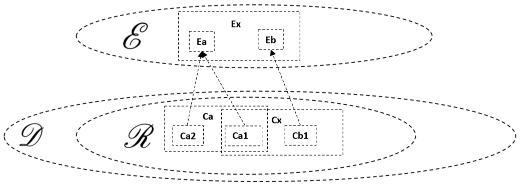


Fig. 1. Space of Capabilities and Entities

techniques ICCF is proposed: a framework for the composition of IoT and CPS capabilities that is based on strong composition foundations provided by the NIST CPS framework, easy and lightweight semantics inspired from the mPlane protocol and leverages the ability to quickly translate service semantics into formal specification thanks to PlusCAL's translation capabilities of composition semantics to TLA specification.

IV. Introducing ICCF composition algebra semantics

This section defines the algebra for describing capabilities, interactions, and compositions.

A. The space of capabilities and descriptors

Figure 1 shows the space of entities and capabilities. $\mathscr{E}$ represents the space of IoT entities, while $\mathscr{D}$ represents the space of capability descriptors. There is a space $\mathscr{R}$ in $\mathscr{D}$ that meets $ICCF$ requirements. $\mathscr{R}$ elements can be composed and decomposed using the ICCF framework specification algebra. There is a $surjective$ relationship between $\mathscr{D}$ and $\mathscr{E}$: one or more $CapabilityDescriptors$ are provided by a single entity ($Ca1$ and $Ca2$ provided by Ea). In the implementation, using microservices, an exception to this rule are those microservices that provide a single and unique $CapabilityDescriptor$ ($Cb1 \rightarrow Eb$ represents this case). If $\mathscr{E}$ is composed of such microservices then the relation between $\mathscr{R}$ and $\mathscr{E}$ is $injective$. Capabilities in $\mathscr{R}$ are either indecomposable or composite.

B. Composition operators and descriptors

Let's define an operator ψ which represents a k-ary composition operator. To illustrate composition in this paper, an assumption of k=2 is considered (which renders ψ a binary composition), Ca1, Ca2, and Ca are represented as JSON objects (with simple key-value pairs representing the $CapabilityDescriptor$ parameters), the composition is an operator on values obtained after sending a specification to all atomic capabilities and receiving results. Let's consider $Ca1$ and $Ca2$ from Figure 1 two indecomposable capabilities and Ca a composite capability obtained as follows:

The composition operator ψ has outcomes in $\mathscr{R}$:

$$\begin{aligned} \psi\texttt{:} \quad & \mathscr{R}^2 \rightarrow \mathscr{R}. \\ \texttt{(Ca1,Ca2)} & \rightarrow \texttt{Ca} \\ & \rightarrow \psi\ \texttt{(Ca1,Ca2)} \end{aligned}$$

This composition generates the $CapabilityDescriptor$ of the composite capability Ca described below:

```
{ "ID": "Ca_ID",
  "Organization": "Ca_O",
  "NAME": "Ca_N",
  "TIMESTAMP": "Ca_TS",
  "LOCATION": "Ca_L",
  "REFRESH_RATE": "Ca_RR",
  "UNIT": "Ca_U",
  "VALUE": "Ca_V",
  "SIGNATURE": "Ca_S"}
```

Ca_ID: represents the ID of the composite capability. It is an increment of the last ID registered in the CMr registry. Ca_N and Ca_O are the Name and the Organization of the new composite capability respectively, a new name and organization are attributed to the composite parameters when the indecomposable capabilities have different ones. Ca_TS: time of arrival of the composite capability. Ca_L represents the physical location (geographical in terms of latitude and longitude) or logical location (IP address). In the case of a geographical address the composite location is the location that comprises indecomposable capabilities' locations. For logical locations, If the sensors reside in the same IP Subnet then the subnet that comprises their IP address becomes their composite location. Ca_RR represents the frequency at which a measurement is received. A composite value for this parameter should be the longest refresh rate:
$Ca\_RR \longleftarrow$ MAX$\{Ca1_RR,Ca2_RR\}$. Ca_U reflects the nature and unit of the composite capability. The simple example of power consumption as a composite capability of both current and voltage takes the "Watt" as the composite Unit of "Amperes" and "Volts". In other cases such as well-being in a smart building, indecomposable capabilities such as temperature, humidity, and air quality have different units and the composition algorithm depicts the composite unit. Ca_V: represents the value of the composite capability. IoT providers have the flexibility to define and introduce parameters customized to their composition needs. One such customization is the introduction of weights and multipliers. For example, $Ca\_V \leftarrow \alpha Ca1\_V + \beta Ca2\_V$ where α and β are two doubles that represent the weight of C_a1V and C_a2V, respectively, and (+) a composition operator. Composition rules and parameters are nested in the programmable extension of the indecomposable capabilities descriptors. This addresses the constructivity concern of the NIST CPS framework as the ability to compose capabilities of different units and sources in a modular way would allow more innovation.

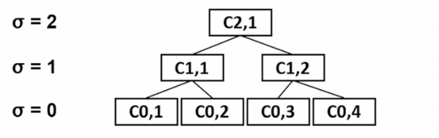


Fig. 2. Composition Hierarchy

C. Capability Hierarchy and Level

Composite capabilities can be further composed into more complex capabilities. $C_{1,1}$ and $C_{1,2}$ in Figure 2 are an example of this case. The hierarchy level σ ranks the capabilities complexity. Every capability can be expressed as follows: $C\sigma, y$, where σ is the hierarchy level and y is the id of the capability at that level. If $\sigma(C)$ =0, the capability is indecomposable. The composite capability descriptor enables tracking the ancestry of the capabilities and verification of their source without directly sending a request to the producing entities. This paradigm addresses the composition complexity concern of the NIST CPS framework.

D. Specifications, results, and interactions

1) Specifications and results: A $Specification$ (sC) for a capability with a descriptor (C) is a request sent from an entity to a resource to get information. The $Specification$ contains information about the capability that helps intermediate entities (including proxies and capability managers) to locate the requested capability. A $Result$ is a $Specification$ for which all the parameters are known. The $Result$ (rC) can be represented as the solution for a system of equations with all the parameters resolved. The space of solutions can contain a unique element, multiple elements, or no element. Below is an example of a $Specification$ represented as a system of equations where the only unknown parameter is C_V: the value of the capability. The other parameters are known as depicted in equation 1. The $Result$ is a unique solution to the $Specification$ as depicted in equation 2.

2) Discovery interaction: The capability manager discovers entities that verify the following rule: $CapabilityDescriptor \in \mathscr{R}$. the discoverability function is defined as Disc(CM, E), it takes CM, a capability manager, and E, an entity as input and returns a binary based on whether or not a capability is discovered. This addresses discoverability, one of the NIST CPS framework composition concerns.

3) Registration interaction: $C \in \mathscr{R} \implies C$ can be registered in CMr. The above implies that all k-ary compositions ψ can be applied to C. The Reg(CM,C) function is a binary function that takes CM (a capability manager) and C (the entity's $CapabilityDescriptor$) as inputs and returns True or false depending on whether or not the descriptor is stored in the CMr and the composition algorithms nested in its programmable extension are stored in the CMt.

$$sC \leftrightarrow \begin{cases} C_{ID} = ID_{value} \\ C_O = O_{value} \\ C_N = N_{value} \\ C_{TS} = TS_{value} \\ C_L = L_{value} \\ C_{RR} = RR_{value} \\ C_U = U_{value} \\ C_V = ? \\ C_S = S_{value} \end{cases} \tag{1}$$

$$rC \leftrightarrow \begin{cases} C_{ID} = ID_{value} \\ C_O = O_{value} \\ C_N = N_{value} \\ C_{TS} = TS_{value} \\ C_L = L_{value} \\ C_{RR} = RR_{value} \\ C_U = U_{value} \\ C_V = V_{value} \\ C_S = S_{value} \end{cases} \tag{2}$$

E. SendSpec(Src,Dst,Specification) interaction

It is a request sC sent to a resource, a proxy, or a capability manager. If C represents a composite capability $CapabilityDescriptor$, the $Specification\ sC$ will be decomposed to its indecomposable $Specifications$ $\big(sC1$,$sC2$,...,$sCk\big)$ first by applying the decomposition operator ψ^{-1} as follows:

$$\begin{array}{rcl} \psi^{-1}\textbf{:} & \mathscr{R} & \rightarrow \ \mathscr{R}^k. \\ & (sC) & \rightarrow \ (sC1, sC2, \ldots, sCk) \\ & & \rightarrow \ \psi^{-1}(sC) \end{array}$$

F. sendResult(Src,Dst,Result) interaction

Entities directly provide a $Result$ for the $Specification$ if it doesn't require further composition or if it is available in the cache CMc. This explains how the adaptability concern of the CPS framework is addressed. Composite $Results$ require composition.

$$\begin{array}{rcl} \psi\textbf{:} & \mathscr{R}^k & \rightarrow \ \mathscr{R}. \\ (rC1, rC2, \ldots, rCk) & & \rightarrow \ rC \\ & & \rightarrow \ \psi(rC1,\ rC2, \ldots, rCk) \end{array}$$

G. $ICCF$ semantics/algebra example

$ICCF$ algebra helps in expressing abstract interactions (discovery, registration, composition, decomposition, specification, results) and enables formal verification, symbolic execution, and making sure the outputs of a system fall within trustworthy values. Pseudo-code in Algorithm 1 summarizes all these operations in a use case explained in the section. The space of capabilities is described above via an example depicted in Figure 1. Let's consider CM, a capability manager, Ex an entity that provides a composite $CapabilityDescriptor\ Cx$ from indecomposable capabilities $Ca1$ and $Cb1$. These indecomposable capabilities are provided by entities Ea and Eb. Ex requests a composite capability Cx from the nearest CM. CM checks its CMc as to whether a copy of the $Result\ rCx$ is available. If this is the case, CM returns the data to to Ex. Otherwise, CM sends requests to entities Ea and Eb based on information in the CMr. These latter respond by sending their $Results$ back to CM. The capability manager performs composition of the $Result$ rCx based on algorithms in the CMt and sends it back to Ex.

Algorithm 1 ICCF Protocol

```
if $Ca1 \in \mathscr{R}$ and $Cb1 \in \mathscr{R}$ then
    Disc($CM$, ($Ea$,$Eb$))← true and
    Reg($CM$,($Ca1$,$Cb1$))← true
    sendSpec($Ex$, $CM$, $sCx$)
if $rCx \in CMc$ then
    sendResult($CM$, $Ex$, $rCx$)
else
    $\psi^{-1}$(sCx)→($sCa1$, $sCb1$)
    sendSpec($CM$, $Ea$, sCa1) and
    sendSpec($CM$, $Eb$, $sCb1$)
    sendResult($Ea$, $CM$, $rCa1$) and
    sendResult($Eb$, $CM$, $rCb1$)
    $\psi$($rCa1$,$rCb1$)→($rCx$)
    sendResult($CM$, $Ex$, $rCx$)
```

So far, ICCF composition interactions and operations were described using semantics inspired from the intuitive mPlane platform and following the capabilities composition guidelines of the NIST CPS framework.

V. Example: Formal specification and assessment of a composite capability: well-being in a smart building

A. Experiment Description

In this section a composition case is studied and its trustworthiness is assessed. It involves composing multiple metrics to get a value-added feature. The example under study is well-being in a smart building: this feature depends on multiple sensor inputs from multiple entities including temperature, humidity, pollution level, and safety sensors in the smart building under study. The well-being composite capability $C5$ is represented as follows:

ψ**:**

$$(rC0, rC1, rC2, rC3, rC4) \rightarrow rC5$$

ψ is the composition operator which represents in this case arithmetic and logical operations on the atomic features $C0$, $C1$, $C2$, $C3$, $C4$ that generate the composite capability $C5$. The goal is to prototype a composition with an assurance, which means the composite feature's values must fall in a trustworthy range. To simplify the specification, the focus is shifted towards the composition of the capabilities data, assuming discovery, registration, and other mPlane protocol interactions are already performed. The composite feature's value, in this case, is a range from 1 to 4 stars representing the level of well-being achieved, with 3 or 4 stars are the trustworthy levels. If security level $rC3$ is not satisfied well-being's value is 0 stars as it is a mandatory aspect. This simple description satisfies the human concern of the NIST CPS framework which guides this implementation. The timestamp is synced across all capabilities values which are discrete. Figure 3 shows the well-being model in PlusCal and its translation to TLA+. Figure 4 shows the range of values generated by each capability and the trustworthy boundaries. Figure 5 shows results after running symbolic execution. The model was run on a Windows Server VM equipped with 4 i9 CPU cores and 24 GB of RAM. TLA+ offers allows connection to remote AWS performant resources to analyze complex and demanding specifications. Figure 6 shows an instance of checking a deadlock state that yields a non-trustworthy outcome.

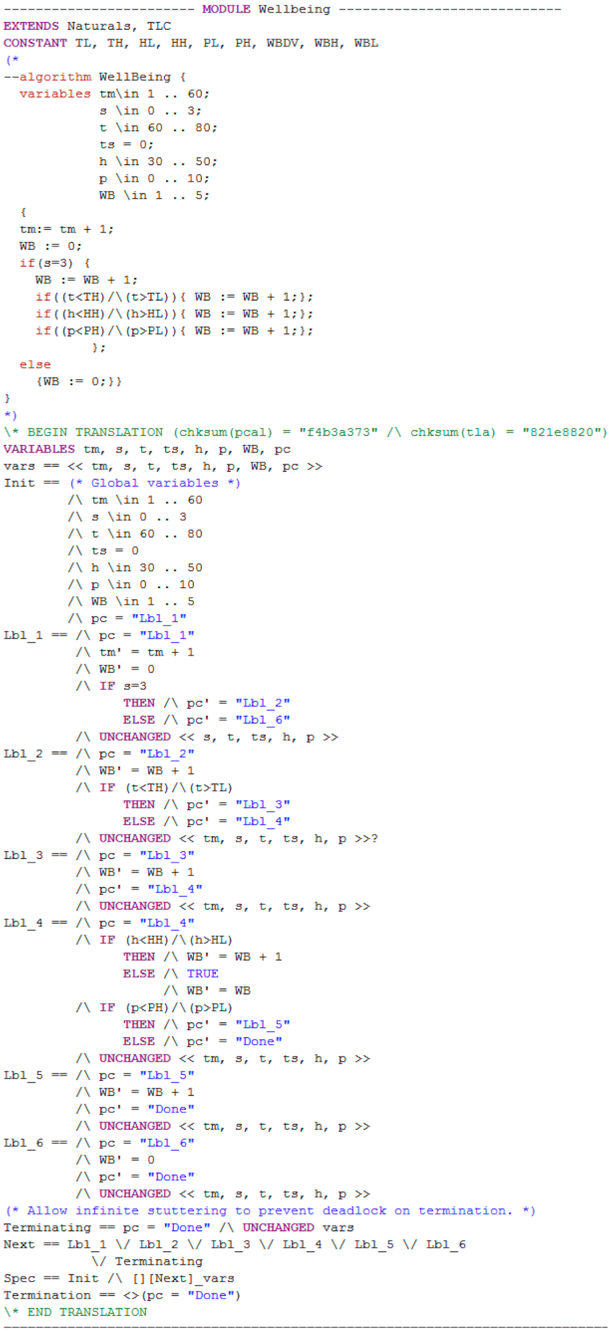
B. Qualitative and Quantitative Analysis

The symbolic execution of the model results in running combinations of all atomic capabilities values to determine the well-being state space. In Figure 4 it took 18 seconds to perform symbolic execution, APALACHE Model Checker can replace TLC to improve execution time [30]. From Figure 5, the number of states generated across all combinations is 14382900, with 5821200 duplicate states, which means more optimization is required. The queue suffered from congestion instantly after execution but it was emptied over. Through this experiment we demonstrated how to prototype a composition based on the framework and semantics proposed, run symbolic execution, analyze trustworthy results, and reveal errors using a deadlock invariant. Other examples that could benefit from this framework include preventing oxygen toxicity in autonomous ventilators (smart health applications) or studying braking time as a composite feature in an autonomous vehicle to evaluate the braking amount required to prevent collisions (smart transportation applications). Minimizing execution time, queue congestion, and the effect of state-space explosion can be done by optimizing the model, space reduction, or leverage of better hardware (local/cloud) which TLA+ allows.

C. Implementation efforts

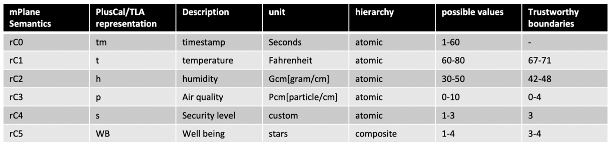
ICCF composition framework is agnostic from the implementation perspective and its principles can be extended to multiple environments. Vert.X, a reactive and event-driven programming tool is used to implement the well-being composite feature based on ICCF

Halba, Khalid; Griffor, Edward; Dahbura, Anton; Lbath, Ahmed. "A Framework for the Composition of IoT and CPS Capabilities." Presented at COMPSAC 2021: Intelligent and Resilient Computing for a Collaborative World 45th Anniversary Conference. The 4th IEEE International Workshop on Smart IoT Sensors & Social Systems for eHealth & Well-Being Applications (SIS-SS 2021). July 12, 2021 - July 16, 2021.



```
------------------------ MODULE Wellbeing ----------------------------
EXTENDS Naturals, TLC
CONSTANT TL, TH, HL, HH, PL, PH, WBDV, WBH, WBL
(*
--algorithm WellBeing {
  variables tm\in 1 .. 60;
            s \in 0 .. 3;
            t \in 60 .. 80;
            ts = 0;
            h \in 30 .. 50;
            p \in 0 .. 10;
            WB \in 1 .. 5;
  {
  tm:= tm + 1;
  WB := 0;
  if(s=3) {
    WB := WB + 1;
    if((t<TH)/\(t>TL)){ WB := WB + 1;};
    if((h<HH)/\(h>HL)){ WB := WB + 1;};
    if((p<PH)/\(p>PL)){ WB := WB + 1;};
           };
  else
    {WB := 0;}}
}
*)
\* BEGIN TRANSLATION (chksum(pcal) = "f4b3a373" /\ chksum(tla) = "821e8820")
VARIABLES tm, s, t, ts, h, p, WB, pc
vars == << tm, s, t, ts, h, p, WB, pc >>
Init == (* Global variables *)
        /\ tm \in 1 .. 60
        /\ s \in 0 .. 3
        /\ t \in 60 .. 80
        /\ ts = 0
        /\ h \in 30 .. 50
        /\ p \in 0 .. 10
        /\ WB \in 1 .. 5
        /\ pc = "Lbl_1"
Lbl_1 == /\ pc = "Lbl_1"
         /\ tm' = tm + 1
         /\ WB' = 0
         /\ IF s=3
               THEN /\ pc' = "Lbl_2"
               ELSE /\ pc' = "Lbl_6"
         /\ UNCHANGED << s, t, ts, h, p >>
Lbl_2 == /\ pc = "Lbl_2"
         /\ WB' = WB + 1
         /\ IF (t<TH)/\(t>TL)
               THEN /\ pc' = "Lbl_3"
               ELSE /\ pc' = "Lbl_4"
         /\ UNCHANGED << tm, s, t, ts, h, p >>?
Lbl_3 == /\ pc = "Lbl_3"
         /\ WB' = WB + 1
         /\ pc' = "Lbl_4"
         /\ UNCHANGED << tm, s, t, ts, h, p >>
Lbl_4 == /\ pc = "Lbl_4"
         /\ IF (h<HH)/\(h>HL)
               THEN /\ WB' = WB + 1
               ELSE /\ TRUE
                    /\ WB' = WB
         /\ IF (p<PH)/\(p>PL)
               THEN /\ pc' = "Lbl_5"
               ELSE /\ pc' = "Done"
         /\ UNCHANGED << tm, s, t, ts, h, p >>
Lbl_5 == /\ pc = "Lbl_5"
         /\ WB' = WB + 1
         /\ pc' = "Done"
         /\ UNCHANGED << tm, s, t, ts, h, p >>
Lbl_6 == /\ pc = "Lbl_6"
         /\ WB' = 0
         /\ pc' = "Done"
         /\ UNCHANGED << tm, s, t, ts, h, p >>
(* Allow infinite stuttering to prevent deadlock on termination. *)
Terminating == pc = "Done" /\ UNCHANGED vars
Next == Lbl_1 \/ Lbl_2 \/ Lbl_3 \/ Lbl_4 \/ Lbl_5 \/ Lbl_6
           \/ Terminating
Spec == Init /\ [][Next]_vars
Termination == <>(pc = "Done")
\* END TRANSLATION
=============================================================================
```

Fig. 3. Well-being model in PlusCal and its translation to TLA



mPlane Semantics	PlusCal/TLA representation	Description	unit	hierarchy	possible values	Trustworthy boundaries
rC0	tm	timestamp	Seconds	atomic	1-60	-
rC1	t	temperature	Fahrenheit	atomic	60-80	67-71
rC2	h	humidity	Gcm[gram/cm]	atomic	30-50	42-48
rC3	p	Air quality	Pcm[particle/cm]	atomic	0-10	0-4
rC4	s	Security level	custom	atomic	1-3	3
rC5	WB	Well being	stars	composite	1-4	3-4

Fig. 4. Capabilities and their Range of possible and accepted values for the well-being composite feature

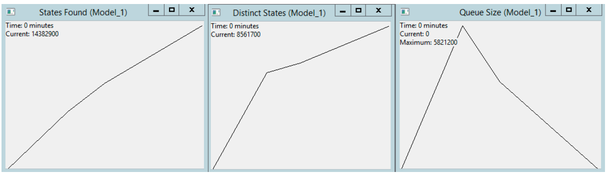
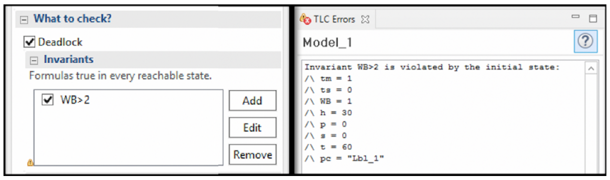


Fig. 5. Symbolic Execution Results as a function of time: States, Distinct States, Queue Size



```
Invariant WB>2 is violated by the initial state:
/\ tm = 1
/\ ts = 0
/\ WB = 1
/\ h = 30
/\ p = 0
/\ s = 0
/\ t = 60
/\ pc = "Lbl_1"
```

Fig. 6. Deadlock and Trustworthiness Verification

foundations. The well-being verticle receives data from temperature and humidity sensors (provided by sensor DHT22 AM2302), and air-quality sensors (provided by sensor SDS011 PM2.5). Code for the project is available in the GitHub repository [31]. An Automated Driving System testbed [32] on the NIST's UCEF co-simulation environment is also being built [33]: The goal is to be able to simulate define autonomy functions as composite features. This will enable trustworthiness assessment of safety critical maneuvers such as emergency braking.

Conclusion and Future work

The ICCF framework and its formal criteria derived from a composition-enabling framework, straightforward and expressive semantics, and strong formal verification language and techniques for composing CPS and IoT capabilities were introduced. A comparison of existing environments, frameworks, semantics, and formal specification and verification techniques enabled the selection of formal components of a composition framework that enables specification, prototyping, and assessment of IoT and CPS capabilities. The goal is to provide stakeholders the tools to innovate in the IoT and CPS space while addressing their corresponding concerns. NIST CPS framework composition guidelines, and powerful semantics inspired from the mPlane protocol, as well as formal specification and verification techniques provided by the TLA/PlusCal package, enable such framework. Composition requirements, services, and interactions were described, and based on that, well-being in a smart building was studied as an example. Results of model checking were generated and an analysis of the state space was performed to understand non-trustworthy results through a deadlock invariant. The following objectives are targeted as future work: strengthening the well-being model as well as tackling the composition concerns in other domains of interest namely: preventing oxygen toxicity in an autonomous ventilator and preventing collisions during emergency braking in the case of autonomous vehicles. Also, as capabilities composition concerns can be mutually exclusive (simplicity vs performance), studying this challenge in-depth is a target milestone.

NIST Disclaimer

Certain commercial equipment, instruments, or materials (or suppliers, or software, ...) 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. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

References

[1] C. Greer, M. J. Burns, D. A. Wollman, and E. R. Griffor, "Cyber-physical systems and internet of things. (no. special publication (nist sp)-1900-202)," tech. rep., 2019.

[2] Cisco, "Cisco iot prediction for 2030. a technical report." https://www.cisco.com/c/en/us/products/collateral/se/internet-of-things/at-a-glance-c45-731471.pdf.

[3] S. Yun, H. Kim, H. Shin, H. S. Chin, and W.-T. Kim, "A novel reference model for cloud manufacturing cps platform based on onem2m standard," *KIPS Transactions on Computer and Communication Systems*, vol. 8, no. 2, pp. 41–56, 2019.

[4] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: Volume 1, overview," 2017.

[5] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: Volume 2, working group reports," 2017.

[6] D. A. Wollman, M. A. Weiss, Y. Li-Baboud, E. R. Griffor, and M. J. Burns, "Framework for cyber-physical systems: Volume 3, timing annex," 2017.

[7] oneM2M Partners, "OneM2M Technical Report : Study of Abstraction and Semantics Enablements." https://onem2m.org/images/files/deliverables/Release2/TR-0018-Industrial_Domain_Enablement-V2_0_0.pdf, 2016.

[8] A. Preventis, K. Stravoskoufos, S. Sotiriadis, and E. G. Petrakis, "Iot-a and fiware: Bridging the barriers between the cloud and iot systems design and implementation.," in *CLOSER (2)*, pp. 146–153, 2016.

[9] A. Bassi, M. Bauer, M. Fiedler, R. van Kranenburg, S. Lange, S. Meissner, and T. Kramp, *Enabling things to talk : Designing IoT solutions with the IoT Architectural Reference Model*. Springer Nature, 2013.

[10] M. Brumbulli, E. Gaudin, and C. Teodorov, "Automatic verification of bpmn models," in *10th European Congress on Embedded Real Time Software and Systems (ERTS 2020)*, 2020.

[11] W. Li, G. Privat, J. M. Cantera, M. Bauer, and F. Le Gall, "Graph-based semantic evolution for context information management platforms," in *2018 Global Internet of Things Summit (GIoTS)*, pp. 1–6, IEEE, 2018.

[12] I. Boneva, J. E. L. Gayo, and E. G. Prud'hommeaux, "Semantics and validation of shapes schemas for rdf," in *International Semantic Web Conference*, pp. 104–120, Springer, 2017.

[13] A. Kazmi, M. Serrano, and J. Soldatos, "Vital-os: An open source iot operating system for smart cities," *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 71–77, 2018.

[14] B. Cheng, G. Solmaz, F. Cirillo, E. Kovacs, K. Terasawa, and A. Kitazawa, "Fogflow: Easy programming of iot services over cloud and edges for smart cities," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 696–707, 2017.

[15] C. Newcombe, T. Rath, F. Zhang, B. Munteanu, M. Brooker, and M. Deardeuff, "How amazon web services uses formal methods," *Communications of the ACM*, vol. 58, no. 4, pp. 66–73, 2015.

[16] L. Lamport, "The pluscal algorithm language," in *International Colloquium on Theoretical Aspects of Computing*, pp. 36–60, Springer, 2009.

[17] L. Lamport, *Introduction to TLA*. Digital Equipment Corporation Systems Research Center [SRC], 1994.

[18] A. Kurniawan, *Learning AWS IoT: Effectively manage connected devices on the AWS cloud using services such as AWS Greengrass, AWS button, predictive analytics and machine learning*. Packt Publishing Ltd, 2018.

[19] B. Trammell, M. Mellia, A. Finamore, S. Traverso, T. Szemethy, B. Szabo, D. Rossi, B. Donnet, F. Invernizzi, and D. Papadimitriou, "mplane architecture specification."

[20] V. Nundloll, Y. Elkhatib, A. Elhabbash, and G. S. Blair, "An ontological framework for opportunistic composition of iot systems," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pp. 614–621, IEEE, 2020.

[21] F. Sivrikaya, N. Ben-Sassi, X.-T. Dang, O. C. Görür, and C. Kuster, "Internet of smart city objects: A distributed framework for service discovery and composition," *IEEE Access*, vol. 7, pp. 14434–14454, 2019.

[22] D. L. McGuinness, F. Van Harmelen, *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[23] G. Simko, D. Lindecker, T. Levendovszky, S. Neema, and J. Sztipanovits, "Specification of cyber-physical components with formal semantics–integration and composition," in *International Conference on Model Driven Engineering Languages and Systems*, pp. 471–487, Springer, 2013.

[24] Y. Li, S. Zhao, H. Diao, and H. Chen, "A formal validation method for trustworthy services composition," in *2016 International Conference on Networking and Network Applications (NaNA)*, pp. 433–437, IEEE, 2016.

[25] C. Laneve and L. Padovani, "An algebraic theory for web service contracts," *Formal Aspects of Computing*, vol. 27, no. 4, pp. 613–640, 2015.

[26] N. Le Sommer, Y. Mahéo, and F. Baklouti, "Multi-strategy dynamic service composition in opportunistic networks," *Information*, vol. 11, no. 4, p. 180, 2020.

[27] H. Groefsema, N. R. van Beest, and M. Aiello, "A formal model for compliance verification of service compositions," *IEEE Transactions on Services Computing*, vol. 11, no. 3, pp. 466–479, 2016.

[28] M. Sozeau, S. Boulier, Y. Forster, N. Tabareau, and T. Winterhalter, "Coq coq correct! verification of type checking and erasure for coq, in coq," *Proceedings of the ACM on Programming Languages*, vol. 4, no. POPL, pp. 1–28, 2019.

[29] T. Ali, M. Nauman, and M. Alam, "An accessible formal specification of the uml and ocl meta-model in isabelle/hol," in *2007 IEEE International Multitopic Conference*, pp. 1–6, IEEE, 2007.

[30] I. Konnov, J. Kukovec, and T.-H. Tran, "Tla+ model checking made symbolic," *Proceedings of the ACM on Programming Languages*, vol. 3, no. OOPSLA, pp. 1–30, 2019.

[31] K. HALBA, "Iotcap : A platform based on the vert.x toolkit and iccf foundations." https://github.com/usnistgov/ICCF.

[32] K. Halba, E. Griffor, P. Kamongi, and T. Roth, "Using statistical methods and co-simulation to evaluate ads-equipped vehicle trustworthiness," in *2019 Electric Vehicles International Conference (EV)*, pp. 1–5, IEEE, 2019.

[33] 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.

Exploring Government Security Awareness Programs: A Mixed-Methods Approach

Jody L. Jacobs, Julie M. Haney, Susanne M. Furman, and Fern Barrientos
National Institute of Standards and Technology

Abstract

Organizational security awareness programs are often underfunded and rely on part-time security awareness professionals who may lack sufficient background, skills, or resources necessary to manage an effective and engaging program. U.S. government organizations, in particular, face challenges due to strict security awareness requirements that often result in success being measured by training completion rates rather than impact on employees' attitudes and behaviors. However, no prior research has explored security awareness in the government sector. To address this gap, we are conducting an in-progress, mixed-methods research effort to understand the needs, challenges, and practices of U.S. government security awareness programs. This understanding will inform the creation of resources for security awareness professionals, including examples of successful practices and strategies, lessons learned, and suggestions for building a team having the appropriate knowledge and skills. While focused on the U.S. government, our findings have implications for organizational security awareness programs in other sectors.

1 Introduction

Despite an abundance of cybersecurity guidance and technologies, organizational employees continue to fall prey to cyber attacks, putting both themselves and their organizations at risk. Security awareness training is a first step towards helping employees recognize and appropriately respond to security issues, with a goal of achieving long-term behavior change [20].

Unfortunately, security awareness efforts face significant challenges. Security awareness programs in organizations of all sizes may be underfunded and rely on part-time security awareness professionals who may lack sufficient background, skills, tools, or resources necessary for managing an effective program [18, 21]. U.S. government – also known as federal – agencies are likewise affected by these challenges as they are mandated to conduct annual security awareness training for all employees [1, 16]. While mandates enforce a minimum baseline for security awareness, when viewed simply as a "check-the-box" exercise, organizations may begin to measure program success simply in terms of compliance metrics, like training completion rates. However, these metrics reveal little about the effectiveness of the training in changing and sustaining workforce attitudes and behaviors [10].

To address the lack of studies about security awareness issues within the U.S. government, we are performing mixed-methods research to better understand the needs, challenges, practices, and necessary competencies of federal security awareness teams and programs. Our research is being conducted in two phases. We held focus groups with federal security awareness professionals to inform the development of a subsequent, online survey that will be sent to a broader population. In this paper, we summarize preliminary results from the focus groups and then briefly describe the planned follow-on survey and potential contributions of our research for both government and non-government organizations.

2 Related Work

Prior research and industry surveys revealed challenges faced by security awareness programs. Programs may receive insufficient attention and funding within their organizations, and security awareness duties are often performed on a part-time, ad-hoc basis [18, 21]. Frequently recruited from the technical security ranks, security awareness professionals may also lack the professional skills (e.g., interpersonal and communication skills) needed to be successful in their role [18].

From a workforce perspective, security awareness training

Workshop on Security Information Workers, USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8, 2021, Vancouver, B.C., Canada.

Jacobs, Jody; Haney, Julie; Furman, Susanne M.; Barrientos, Fernando. "Exploring Government Security Awareness Programs: A Mixed-Methods Approach." Presented at Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021). August 08, 2021 - August 10, 2021.

may be viewed as an inconvenient, boring, "check-the-box" exercise with little relevance to day-to-day work [5, 11]. To counter these challenges, researchers [2, 4, 5, 7] recommend that programs better engage employees by communicating how security impacts the organization, tailoring communications to various audiences, and implementing creative ways to disseminate awareness information. In addition, programs should continuously provide training refreshers throughout the year to help make security a habit both at work and home.

Measuring program success is an important but often overlooked aspect of security awareness programs. For a holistic assessment, recommendations point to organizations using a combination of measures, such as security incident trends and reporting, views/engagement with security awareness materials, and feedback from stakeholders [4, 10].

Although evidence of security awareness challenges and recommendations abound, it is currently unknown whether these apply to programs within the U.S. government sector and if government organizations experience additional issues. Our research addresses this gap.

3 Methodology

We are undertaking an exploratory sequential mixed methods research approach (qual → QUAN) [8], with focus groups informing a broader survey. In this paper, we describe the completed focus group study phase and briefly describe our future plans for the survey.

3.1 Study Design

Focus groups can be valuable when used as a precursor to quantitative surveys of larger samples as they can facilitate the development of survey questions by providing an understanding of how people talk about specific topics and what concepts are most important [12, 14]. We selected focus groups, rather than interviews, for several reasons. Since one of the goals of our study was to identify potential ways in which information could be shared more effectively across the community, it was valuable to observe how ideas emerged during group discussion, which is not possible in individual interviews. Focus groups also served a practical purpose as we had an abbreviated timeline in which to collect and analyze data. We hoped that our study results could inform the revision of a federal security awareness guidance document [20] set to commence around the same time as our study. Wanting to provide input earlier rather than later in the revision process and factoring in the time to design and execute a follow-on survey, focus groups were deemed a more efficient way to collect data as compared to individual interviews.

When designing the focus group study, we consulted seven federal security awareness subject matter experts (SMEs). The final protocol consisted of 11 questions covering topics such as approaches, successes, challenges, measuring effectiveness, wish lists, and necessary knowledge and skills for security awareness teams. We selected a multiple-category design, which involves focus groups with several types of participants to allow for comparisons across or within categories [12]. Based on SME discussions, we decided on three categories: 1) department-level organizations (e.g., U.S. Department of Labor), 2) sub-component agencies, which are semi-autonomous organizations under a department (e.g., Bureau of Labor Statistics under Department of Labor), and 3) independent agencies, which are not in a department. In the Executive Branch of the U.S. Government, there are 15 departments, over 200 sub-components, and just over 100 independent agencies.

Potential focus group participants were selected via a purposeful approach to identify information-rich cases and represent diversity of agencies. We identified participants via several avenues: recommendations from the SMEs; researchers' professional contacts; a mailing list of small and micro agencies; previous speakers and participants from the last three years of the Federal Information Security Educators (FISSEA) conference [15]; and LinkedIn and Google searches. Participants had to have knowledge of the security awareness programs in their organizations either because they had security awareness duties or oversaw the programs.

3.2 Data Collection

Between December 2020 and January 2021, we conducted eight virtual focus groups with 29 total participants: 2 groups with 6 department-level participants, 3 with 12 participants from independent agencies, and 3 with 11 participants from sub-components. Group sessions lasted 60-75 minutes, with each having 3-5 participants. We found that, given the virtual nature of the groups, smaller numbers of participants worked best [19]. All groups were audio recorded and transcribed. Participants also completed an online survey to gather demographic and organizational information.

The study was approved by our institution's Research Protections Office with informed consent required for all participants. To ensure anonymity, each participant was assigned a reference code, with individuals from independent agencies identified as N01 – N12, department-level organizations as D01 – D06, and sub-components as S01 – S11.

3.3 Data Analysis

For data analysis, initially, each member of the four-person research team individually coded a subset of three transcripts (one from each category) using an *a priori* code list based on research questions and open coded for additional concepts as needed. We met several times to discuss codes and develop a codebook. In accordance with the recommendation of qualitative methodologists [6, 13], we focused not on calculating agreement scores but rather on how and why disagreements in coding arose and the insights afforded by subsequent discussions. When disagreement occurred, we discussed as a group to reach consensus.

Coding continued until each remaining transcript was coded by two researchers. The coding pair then met to discuss code application and resolve differences. The entire research team convened to discuss overarching themes identified in the data.

4 Participant Demographics

The 29 participants represented 28 unique federal organizations (one agency had two people participate). Among those, 22 led their respective security awareness programs, three were security awareness team members, and four were managers or Chief Information Security Officers (CISOs). All but two were part-time in their security awareness duties (average 46%). Twenty had been involved in security awareness for more than five years, with the others involved 1-5 years. Eleven of the 23 who provided formal education information had at least one degree in a technology-related field.

Fifteen participants were male and 13 female, with one participant not disclosing gender. One participant was in the 18-29 years-old age range, 5 were 30-39, 9 40-49, 9 50-59, and two 60 + (3 did not disclose age).

5 Preliminary Results

5.1 Required Annual Training

U.S. government organizations are mandated to implement annual, mandatory security awareness training programs for their workforce. In executing these programs, security awareness professionals encounter several challenges. Employees may perceive the training as a "check-the-box" exercise with boring, unchanging content and are often overwhelmed by having to complete numerous other mandatory organizational training courses. As one participant mentioned, "You've got IT security, physical security, personnel security, etc. And they have their own training requirements. . . So to me, it's inefficient for our user base not to have one course that meets all the needs" (S11). In addition, programs often have challenges tracking training completion of contractors supporting the organization or seasonal staff due to them not having the same system access as government employees.

We noted that enforcement of awareness training completion varied among organizations. Nine participants indicated that their organization took a zero-tolerance approach by disabling accounts of employees who failed to complete the training by the appointed deadline. One participant described how enforcement, while having its merits, also resulted in additional challenges for his organization: "They [employees] put it off. . . Even though we're giving them messages throughout the year, they'll wait. And then when we had to come up with this big, long list of people we're disabling accounts, then it becomes a political nightmare" (N08). Conversely, five participants expressed that they had not received the organizational support necessary to enforce training completion, especially when leaders themselves are guilty of not completing the training: "Our biggest problem is with our executives. They are the ones who are more than likely not to have taken the training in a timely manner, and we can't exactly lock them out" (S03).

5.2 Approaches

The programs represented in our focus groups delivered their required, annual cybersecurity training via standard online, computer-based or instructor-led training. Furthermore, most organizations went above-and-beyond mandatory training and disseminated security awareness information throughout the year via a variety of methods, including newsletters, cybersecurity tips of the month, broadcast emails, posters, speaker events, and webinars. Phishing simulation exercises – in which employees are sent emails that mimic real-world phishing attempts to train them to recognize and appropriately respond to phishing emails – were particularly popular. Several organizations took novel approaches to deliver information – for example, escape rooms and virtual reality – with the intent of boosting employee engagement. Programs also hold awareness campaigns in line with annual National Cybersecurity Awareness Month themes [9].

With the variety of security awareness delivery methods, participants noted that they were highly sensitive to the amount, relevancy, and conciseness of information they distributed to their workforce as they did not want to overwhelm employees. Participants expressed challenges in ensuring their delivery methods are adaptable to a variety of learning styles, skill levels, and work roles. They also discussed difficulties they face in meeting various accessibility and assistive technology requirements, especially when implementing novel techniques such as virtual reality.

When asked what is working well for their programs, three participants mentioned that they have incentive programs to engage users and reward them for good security behaviors. One organization incorporated a multi-level gamification approach where individuals could advance as their security awareness increased. Another gave employees a certificate to hang in their office or added a "badge" in their email signature stating they were a "phish hunter" when they successfully reported a phish during a scheduled phishing exercise. These incentive programs resulted in "a lot of internal, healthy competition" (N01) and encouraged engagement with security awareness information.

5.3 Security Awareness Content

We found that organizations have no single resource from which to obtain security awareness training materials. Nineteen participants indicated that their organization outsourced at least some content development to an external entity (e.g., contractor, training vendor, or other federal organization), while others developed content in-house. Eight participants from sub-component organizations stated that they received complete or partial training materials from their parent department, with most having the autonomy to customize the training to fit their organizations' unique needs. Programs determined which security topics to cover in a variety of ways, often utilizing external sources (e.g., SANS, national news outlets) or internal sources (e.g., their organization's security operations center, workforce feedback) to identify pertinent security topics and trends. Participants felt that their workforce responded positively when training topics were relevant and relatable to both the organization and employees' daily lives.

When asked what resources might best help their programs, participants offered suggestions. Because finding or developing awareness content was viewed as a challenge, 15 participants expressed that security awareness programs would benefit from having a single, federal-level security awareness course to fulfill mandatory training requirements. The training would include core materials common to all organizations while allowing programs to customize some content for their current environment and organizational mission. A standardized course would ensure the delivery of consistent security awareness information and reduce the burden on federal security awareness programs. A participant who supported this idea stated, "There are... probably 80% of the topics everybody needs to know about. So, why are we buying that over and over again at each agency as opposed to give us the 80% solution and let us pay for the other 20%? That would be more efficient" (D01).

Other participants suggested a repository where organizations could share awareness materials to augment other programs' offerings. One participant said:

> "if there was a central repository within the federal government... of various trainings and awareness pamphlets, flyers, presentations... that the various agencies could actually share and leverage back and forth, I think that would definitely better help us make use of what limited resources we do have" (S01).

Participants also desired to have more government-specific, detailed guidance regarding security awareness training content/topics and delivery methods and tools. Without clear direction, many programs have had to interpret federal polices and directives on their own, leading to marked differences in training quality across organizations. As one participant said, "I think that's something that we could use more guidance on. How long does the course have to be? Does it have to be specific?...We've asked for that guidance on a consistent basis, but all we have is the general guidance" (S04).

Only five participants noted that they are involved with security awareness working groups and online forums where the latest security developments and training approaches are shared. To encourage more cross-organizational collaboration, participants suggested that a real-time sharing platform for federal security awareness professionals would be beneficial. This platform would create an environment where lessons learned, trend analysis, training opportunities, and approaches could be shared. One participant stated, "if we... share the results, we can help each other build more efficient programs for our respective agencies" (D02).

5.4 Measures of Effectiveness

Participants employed a variety of methods to determine the effectiveness of their security awareness programs. Training completion rates were a popular metric, but some participants acknowledged that these do not demonstrate long-term attitude or behavior change, which should be the real goals of security awareness training. Participants mentioned other indicators of success, such as: security awareness event attendance; employee feedback, including formal (e.g., via surveys) and informal (e.g., via personal interactions); and program audit reports. Six participants indicated that they review user-generated security incidents, security operations trends, and reporting to help determine whether certain security topics are being effectively taught and translated into action by the workforce. Participants also use click and reporting rates collected during their phishing simulations to determine the effectiveness of phishing-related training.

Although all programs make at least some attempt, 13 participants are still unsure how to gauge effectiveness. As one participant said, "We run security awareness campaigns and... we really have no idea how much of it is absorbed" (S04). Participants expressed a desire for more government guidelines on ways to measure effectiveness. One security awareness professional spoke of the benefits of standardized measures that could help "determine whether or not the programs that are out there are effective or what parts need to actually be focused on" (S01).

5.5 Team Knowledge and Skills

We asked what knowledge and skills a security awareness professional should possess. Sixteen participants stated that technical knowledge was highly important, but others felt that this knowledge could be outsourced to other security staff in the organization. Additionally, non-technical, professional skills such as interpersonal, communication, creativity, and collaboration were mentioned as being just as, if not more, important as a technical background.

Participants agreed that finding a single individual who possesses all desired knowledge and skills would be ideal, but may be difficult to achieve. Therefore, building a multidisciplinary team can be beneficial. As one participant shared, "I have people who can design, are very artful, creative people. I have people who can run a learning management system…I have good project managers. I have cybersecurity professionals" (D01). However, some programs do not have the resources for an entire team and instead must rely on one person to run the entire program. In these cases, it becomes especially important to work closely with other components of the organization (e.g., human resources, communications, and the training group) to assist with activities such as outreach and training material development.

6 Future Work and Implications

Our analysis of the focus group data identified areas of interest that will inform the development of an online survey to be sent to a larger population of federal security awareness professionals. We will synthesize the qualitative focus group data with the largely quantitative survey data to capture a deeper understanding of the state, challenges, and experiences of U.S. government security awareness programs.

It is our hope that insights gained from our research will lead to the creation of multiple resources for federal security awareness training professionals, including: examples of successful practices and strategies; lessons learned; suggestions for building a team with appropriate core competencies; and the creation of information sharing platforms, such as an online forum, working group, or central repository. In addition, results will inform government-wide guidelines to aid federal organizations in the development of effective security awareness training programs.

Even though we are focusing on federal security awareness programs in the U.S., our findings appear to have relevance to programs in other countries', e.g., [3, 17].

In addition, there are other sectors outside the government that implement security awareness training and are mandated to do so, like the health and financial communities. Therefore, we believe many of our findings may be transferable to non-federal organizations.

Disclaimer

Any mention of commercial products or companies is for information only and does not 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 purpose.

References

[1] 113th Congress. Federal Information Security Modernization Act of 2014 (FISMA). https://www.congress.gov/bill/113th-congress/senate-bill/2521/text, 2014.

[2] Jemal Abawajy. User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, 33(3):237–248, 2014.

[3] Raneem AlMindeel and Jorge Tiago Martins. Information security awareness in a developing country context: insights from the government sector in Saudi Arabia. *Information Technology & People*, May 2020.

[4] Moneer Alshaikh, Sean B. Maynard, Atif Ahmad, and Shanton Chang. An exploratory study of current information security training and awareness practices in organizations. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, January 2018.

[5] Maria Bada, Angela M. Sasse, and Jason R.C. Nurse. Cyber security awareness campaigns: Why do they fail to change behaviour? https://arxiv.org/ftp/arxiv/papers/1901/1901.02672.pdf, 2019.

[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] Stefan Bauer, Edward W. Bernroider, and Katharina Chudzikowski. Prevention is better than cure! Designing information security awareness programs to overcome users' non-compliance with information security policies in banks. *Computers & Security*, 68:145–159, 2017.

[8] Vicki L. Plano Clark. Meaningful integration within mixed methods studies: Identifying why, what, when, and how. *Contemporary Educational Psychology*, 57:106–111, 2019.

[9] Cybersecurity and Infrastructure Security Agency. National Cybersecurity Awareness Month (NCSAM). https://www.cisa.gov/national-cyber-security-awareness-month, 2021.

[10] Tobias Fertig, Andreas E. Schütz, and Kristen Weber. Current issues of metrics for information security awareness. In *Proceedings of the European Conference on Information Systems*, 2020.

[11] Julie Haney and Wayne Lutters. Security awareness training for the workforce: Moving beyond "check-the-box" compliance. *Computer*, 53(10):91–95, 2020.

[12] Richard A. Krueger and Mary Anne Casey. *Focus Groups: A Practical Guide for Applied Research*. Sage, Thousand Oaks, CA, 5th edition, 2015.

[13] 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.

[14] Sylvia C. Nassar-McMillan and L. DiAnne Borders. Use of focus groups in survey item development. *The Qualitative Report*, 7(1):1–12, 2002.

[15] National Institute of Standards and Technology. FISSEA – Federal Information Security Educators. https://csrc.nist.gov/projects/fissea, 2021.

[16] Office of Management and Budget. Circular A-130. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A130/a130revised.pdf, 2016.

[17] Eka Ayu Puspitaningrum, Ferizka Tiara Devani, Vidya Qoriah Putri, Achmad Nizar Hidayanto, and Ika Chandra Hapsari. Measurement of employee information security awareness: Case study at a government institution. In *Proceedings of the 2018 Third International Conference on Informatics and Computing (ICIC)*, pages 1–6, 2018.

[18] SANS. 2021 SANS security awareness report: Managing human cyber risk. https://www.sans.org/security-awareness-training/resources/reports/sareport-2021/, 2021.

[19] UXalliance. Conducting remote online focus groups in times of COVID-19. https://medium.com/@UXalliance/conducting-remote-online-focus-groups-in-times-of-covid-19-ee1c66644fdb, April 2020.

[20] Mark Wilson and Joan Hash. NIST Special Publication 800-50 - Building an information technology security awareness program. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-50.pdf, 2003.

[21] Ben Woelk. The successful security awareness professional: Foundational skills and continuing education strategies. https://library.educause.edu/~/media/files/library/2016/8/erb1608.pdf, 2015.

REALTIME UNCERTAINTY QUANTIFICATION VIA ULTRA-PRECISE PARTICLE MATCHING FOR HIGH-THROUGHPUT SERIAL CYTOMETRY

Matthew DiSalvo[1,2], Paul N. Patrone[2], and Gregory A. Cooksey[2]
[1]Johns Hopkins University, USA and
[2]National Institute of Standards and Technology, USA

ABSTRACT

A central paradigm in flow cytometry is the one-time optical interrogation of cells, a practice that has limited the ability to address important questions associated with reproducibility and repeatability of measurements. Serial flow cytometry has pioneered the quantification of measurement uncertainties by optically interrogating each object more than once along a flow path. Here, we address the throughput limitations of serial cytometry with an algorithm to match signals across different interrogation regions. The algorithm operated real-time in an automated microfluidic serial cytometer and matched 99.96 % (95 % confidence interval [99.91 %, 99.98 %]) of particles at 94 Hz.

KEYWORDS: Flow Cytometry, Microfluidics, Optofluidics, Reproducibility

INTRODUCTION

Previously, we reported on a microfluidic serial cytometer that used integrated waveguides and a novel inertial and 3-D hydrodynamic flow focusing strategy to demonstrate the feasibility of quantifying measurement uncertainties in individual events [1, 2]. The cytometer achieved particle velocity variations of ≈ 0.3 % and median fluorescence area measurement precisions of ≈ 2 % from calibration microspheres. However, operation of the device was restricted to a throughput of 1 Hz to avoid challenges associated with matching, thus limiting its utility.

THEORY

In a strict approach to particle matching, unambiguous matching proceeds if and only if the time-of-flight (TOF) is less than interparticle latency (Figure 1). In contrast, for forward-projection time subdivision (FPTS), a particle was matched if, for the k^{th} non-reference signal channel, there existed exactly 1 index m_k where the peak time $t_{m_k,k}$ landed between two sequential time boundaries. The boundaries $b_{n,k}$, dividing the time for $n = 1 \ldots N$ particles shifted by estimated TOFs $\delta_{n,k}$ and peak times $t_{n,ref}$ from the reference channel, are given by Equation (1). Otherwise, series of particles whose signals did not uniquely occupy an equally-sized series of time windows were matched in order.

$$b_{n,k} = \begin{cases} t_{n,ref} + \frac{t_{n+1,ref} - t_{n,ref}}{2} + \delta_{n,k}, & n < N \\ \infty, & n = N \end{cases} \tag{1}$$

EXPERIMENTAL

An average of 10 000 green fluorescent polystyrene microspheres (15.3 µm nominal diameter) were measured under each of 19 different flow-focusing conditions by a microfluidic serial flow cytometer configured with six detectors (two fluorescence and one transmission for each of two laser excitation regions). The conditions included particle-based Reynold's numbers ranging from 2.7 to 3.7, sheath-to-core ratios (SCRs) from 1 to 130, and event rates from 2 Hz to 760 Hz. Simulated events were timed according to a Poisson process with invariant event order. Confidence intervals (CIs) for matching proportions represent Clopper-Pearson binomial proportion intervals; CIs for velocities represent bias-corrected and accelerated bootstrap intervals.

RESULTS AND DISCUSSION

The performances of the strict and FPTS algorithms were evaluated on experimental and synthetic signals from particles at various velocities, velocity variations, and event rates (Figure 2). The tracking yield of both algorithms were limited by event rates; however, while the strict approach improved at higher velocities, the FPTS algorithm improved at lower velocity variations. At ideal operating conditions, experimental and synthetic data indicate that the failure point (loss > 0.1 %) occurs at ≈ 1 Hz for the strict approach and ≈ 100 Hz for FPTS. The FPTS algorithm was incorporated into automation routines and particles flowing at 0.771 ± 0.020 m/s (mean ± standard deviation of 27 461 particles at 762 Hz) were successfully analyzed, displayed, and logged in real-time (4 s buffer and 500 ns sampling interval). Novel metrics afforded by serial nature of the measurements, such as particle velocity and measurement precision, were used to support on-the-fly performance tracking of the instrument under various flow focusing conditions (Figure 3). When challenged with a 556 Hz event rate and low hydrodynamic flow focusing

DiSalvo, Matthew; Patrone, Paul; Cooksey, Gregory A. "Realtime Uncertainty Quantification Via Ultra-Precise Particle Matching for High-Throughput Serial Cytometry." Presented at 25th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS 2021), Palm Springs, CA, US. October 10, 2021 - October 14, 2021.

(SCR = 3), the matching yield was 98.96 % (95 % CI [98.82 %, 99.10 %]) and the precision of integrated fluorescence area was 0.7 % to 1.8 % (25th to 75th percentile of individual particle replication coefficient of variations).

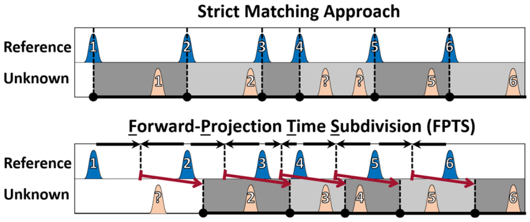


Figure 1: Particle matching strategies. Top: Signals in a reference channel (blue) are matched to downstream channel (tan) with unknown peak identities only if one unknown peak uniquely occupied the time windows (grey) spanning reference signals. Bottom: In FPTS, time window boundaries are extrapolated away from halfway (black arrows) between reference peaks using an estimated time-of-flight (red arrows).

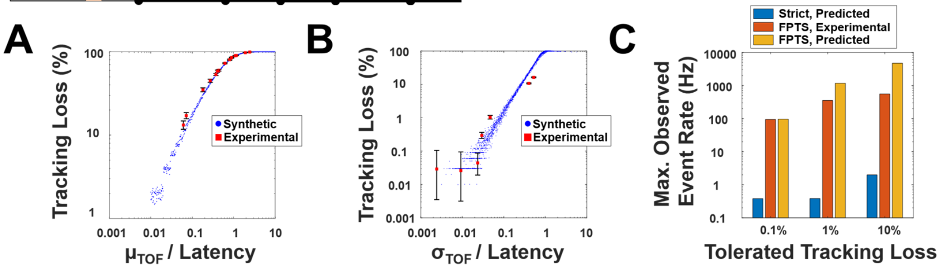


Figure 2: Characterization of particle matching limits. A) Strict matching and B) FPTS algorithm. Error bars represent 95 % binomial proportion confidence intervals. μ_{TOF} : mean TOF; σ_{TOF} : sample standard deviation of TOF; latency : time between sequential particles; tracking loss : proportion of particles detected but not matched across replicate measurement channels . Each synthetic data point is the result from N = 10 000 particles. C) Maximum observed event rates across synthetic and experimental datasets with tracking loss (95 % binomial proportion confidence interval) below tolerances. Predictions assumed ideal operating conditions of $\mu_{TOF} \approx 26$ ms and $\sigma_{TOF}/\mu_{TOF} \approx 0.2$ %. No experimental data had less than 10 % tracking loss using the strict matching method; the lowest observed loss using strict matching was 13 % at 2.4 Hz.

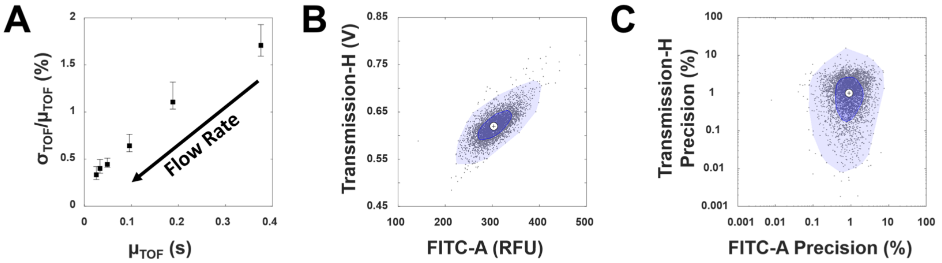


Figure 3: Serial cytometry results. A) Monitoring of particle velocity and velocity variability with changing flow rates. Error bars represent 95 % bootstrap confidence intervals. B) Traditional flow cytometry scatterplot of forward scatter versus fluorescence for N = 4305 particles with 100 % matching yield acquired at 35 Hz. C) Scatterplot of the same particles with the novel axes of measurement precisions. B,C) Dark blue: envelope containing central 50 % of data points; Light blue : envelope containing inliers; White crosshair : Tukey median. Transmission-H : height of the signal representing loss of transmitted light due to particle crossing the laser path (analogous to forward scatter); FITC-A : integrated area of the signal representing green fluorescence intensity emitted by the particles.

CONCLUSION

Automated particle matching enabled on-the-fly uncertainty quantification of flow cytometry measurement reproducibility *on a per-event basis*, alongside continual and simultaneous measurement and data logging. The approach allows us to characterize when the measurement system becomes unstable with respect to measurement reproducibility. We anticipate that serial cytometry will reveal and quantify additional sources of uncertainty arising from the instrumentation, sample, and analyses and provide better tools to compare rare events within a population.

REFERENCES

[1] M. DiSalvo, P. N. Patrone, G. A. Cooksey, in *24th Int. Conf. Miniaturized Syst. Chem. Life Sci*, 835, Oct. 2020.
[2] G. A. Cooksey, P. N. Patrone, J. R. Hands, *et al.*, *Anal. Chem.*, 91, 16, 10713–10722, Aug. 2019

CONTACT: G.A. Cooksey; phone: +1-301-975-5529; gregory.cooksey@nist.gov

DiSalvo, Matthew; Patrone, Paul; Cooksey, Gregory A. "Realtime Uncertainty Quantification Via Ultra-Precise Particle Matching for High-Throughput Serial Cytometry." Presented at 25th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS 2021), Palm Springs, CA, US. October 10, 2021 - October 14, 2021.