NIST Special Publication 800
NIST SP 800-239 ipd

AI Data Center Security Analysis

A High-Performance Computing (HPC) Driven Approach

Yang Guo
Bennett Tomlinson

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-239.ipd

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

NIST Special Publication 800
NIST SP 800-239 ipd

AI Data Center Security Analysis

A High-Performance Computing (HPC) Driven Approach

Yang Guo
Computer Security Division
Information Technology Laboratory

Bennett Tomlinson
Center for AI Standards and Innovation

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-239.ipd

July 2026

U.S. Department of Commerce
Howard Lutnick, Secretary of Commerce

National Institute of Standards and Technology
Arvind Raman, NIST Director and Under Secretary of Commerce for Standards and Technology

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 (NIST), nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. Many NIST cybersecurity publications, other than the ones noted above, are available at https://csrc.nist.gov/publications.

Authority
This publication has been developed by NIST in accordance with its statutory responsibilities under the Federal Information Security Modernization Act (FISMA) of 2014, 44 U.S.C. § 3551 et seq., Public Law (P.L.) 113-283. NIST is responsible for developing information security standards and guidelines, including minimum requirements for federal information systems, but such standards and guidelines shall not apply to national security systems without the express approval of appropriate federal officials exercising policy authority over such systems. This guideline is consistent with the requirements of the Office of Management and Budget (OMB) Circular A-130.

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on federal agencies by the Secretary of Commerce under statutory authority. Nor should these guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce, Director of the OMB, or any other federal official. This publication may be used by nongovernmental organizations on a voluntary basis and is not subject to copyright in the United States. Attribution would, however, be appreciated by NIST.

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on [Will add date upon final publication].

How to Cite this NIST Technical Series Publication:
Guo Y, Tomlinson B (2026) AI Data Center Security Analysis: A High-Performance Computing (HPC) Driven Approach. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) NIST SP 800-239 ipd. https://doi.org/10.6028/NIST.SP.800-239.ipd

Author ORCID iDs
Yang Guo: 0000-0002-3245-3069
Bennett Tomlinson: 0009-0005-6434-5758

Public Comment Period
July 27, 2026 – September 25, 2026

Submit Comments
sp800-239-comments@nist.gov

National Institute of Standards and Technology
Attn: Computer Security Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8930) Gaithersburg, MD 20899-8930

Additional Information
Additional information about this publication is available at https://csrc.nist.gov/pubs/sp/800/239/ipd, including related content, potential updates, and document history.

All comments are subject to release under the Freedom of Information Act (FOIA).

Abstract

The architecture of high-performance computing (HPC) systems has significantly influenced the development of artificial intelligence (AI) data centers, which are purpose-built for model training, inference, and applications. This publication conducts a threat and security gap analysis of AI data centers and provides basic recommendations. It leverages prior work on HPC threat analysis, security posture, and the HPC security overlay to identify key differences between AI data centers and HPC systems in system architecture, hardware, software stacks, workflows, and data storage systems. The key threats facing AI data centers are then identified, and possible solutions are included.

Keywords

AI data center; AI data center security; AI data center computing environment; AI data center reference architecture; high-performance computing (HPC); HPC reference architecture; HPC security; security guideline.

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems. The Special Publication 800-series reports on ITL's research, guidelines, and outreach efforts in information system security, and its collaborative activities with industry, government, and academic organizations.

Note to Readers

This publication is a joint effort between the Information Technology Laboratory (ITL) and the Center for AI Standards and Innovation (CAISI). Uniting both groups' extensive expertise in cybersecurity, artificial intelligence, and infrastructure resilience, this collaboration establishes vital research and technical guidelines to support the security of AI data centers. As high-performance AI workloads scale rapidly, this unified approach is essential for delivering robust guidelines that help secure critical computing infrastructure against emerging risks.

Call for Patent Claims

This public review includes a call for information on essential patent claims (claims whose use would be required for compliance with the guidance or requirements in this Information Technology Laboratory (ITL) draft publication). Such guidance and/or requirements may be directly stated in this ITL Publication or by reference to another publication. This call also includes disclosure, where known, of the existence of pending U.S. or foreign patent applications relating to this ITL draft publication and of any relevant unexpired U.S. or foreign patents.

ITL may require from the patent holder, or a party authorized to make assurances on its behalf, in written or electronic form, either:

a) assurance in the form of a general disclaimer to the effect that such party does not hold and does not currently intend holding any essential patent claim(s); or

b) assurance that a license to such essential patent claim(s) will be made available to applicants desiring to utilize the license for the purpose of complying with the guidance or requirements in this ITL draft publication either:

   i. under reasonable terms and conditions that are demonstrably free of any unfair discrimination; or

   ii. without compensation and under reasonable terms and conditions that are demonstrably free of any unfair discrimination.

Such assurance shall indicate that the patent holder (or third party authorized to make assurances on its behalf) will include in any documents transferring ownership of patents subject to the assurance, provisions sufficient to ensure that the commitments in the assurance are binding on the transferee, and that the transferee will similarly include appropriate provisions in the event of future transfers with the goal of binding each successor-in-interest.

The assurance shall also indicate that it is intended to be binding on successors-in-interest regardless of whether such provisions are included in the relevant transfer documents.

Such statements should be addressed to: sp800-239-comments@nist.gov

Table of Contents

List of Figures

1. Introduction

Artificial intelligence (AI) is a transformative technology whose far-reaching applications are creating new industries and fundamentally changing how we live and work. As a result, AI data centers have experienced significant growth in capability, scale, and investment. However, these data centers also face considerable challenges, including power consumption, intellectual property protection, and AI model security. Ensuring the security of these data centers is crucial for achieving AI security and trustworthiness.

The AI Action Plan [1] calls for establishing new technical standards for high-security AI data centers, which are specifically designed to train AI models and perform AI inference and other services that often demand significant computing power, efficient parallel processing, and the ability to manage and process large volumes of data.

This publication focuses on the security of the computing environment in AI data centers, specifically addressing access, management, computation, and data storage for AI workloads. This computing environment is one aspect of a complete AI data center. Comprehensive security for an AI data center must also consider other elements, such as the site perimeter, the building and physical infrastructure, the facility's operational technology (OT) that manages power and cooling, and the supply chain and personnel involved in operations. These components are beyond the scope of this document and will be covered by a correlative initiative.

The development of a security framework for the entire AI data center is an ongoing effort. The threat analysis and recommendations in this publication shall lay the groundwork for creating security guidelines specific to the computing environment. These guidelines will likely become an essential part of the overall security framework for AI data centers. For the remainder of this document, we will use the terms "AI data center computing environment" and "AI data center" interchangeably.

An AI data center bears many similarities to a high-performance computing (HPC) system. High-performance computing (HPC) systems are designed to serve as the computing infrastructure for scientific exploration, including complex simulations and the resolution of partial differential equations (PDEs). HPC is becoming increasingly important for AI training and big data analysis [2] due to its ability to support high-performance parallel computing. HPC architectures have significantly influenced the development of today’s AI data center.

1.1. Zone-Based HPC Architecture

In NIST SP 800-223 [3], a zone-based HPC reference architecture is proposed as shown in Figure 1. An HPC system is structured in four functional zones:

1. The *computing zone* (i.e., high-performance computing zone) consists of a pool of compute nodes that are connected by one or more high-speed networks. The high-performance computing zone provides key services specifically designed to run parallel jobs at scale.

1. The *data storage zone* comprises layered data storage systems with one or more high-speed, parallel file systems (PFSs) that are designed to store very large datasets and provide fast read and write access.
2. The *access zone* provides connectivity services to external networks, such as the broader organizational network or the internet. This zone authenticates and authorizes user and administrator access. It also provides a variety of services, such as interactive shells, web-based portals, data transfer, and data visualization.
3. The *management zone* comprises multiple management nodes and/or cloud service clusters that provide HPC management services. This zone allows HPC system administrators to configure and manage the HPC system, including the compute node configuration, storage and network provisioning, identity management, auditing, system monitoring, and vulnerability assessment. Various management software modules (e.g., job schedulers, workflow management, Domain Name System [DNS]) also run in the management zone.

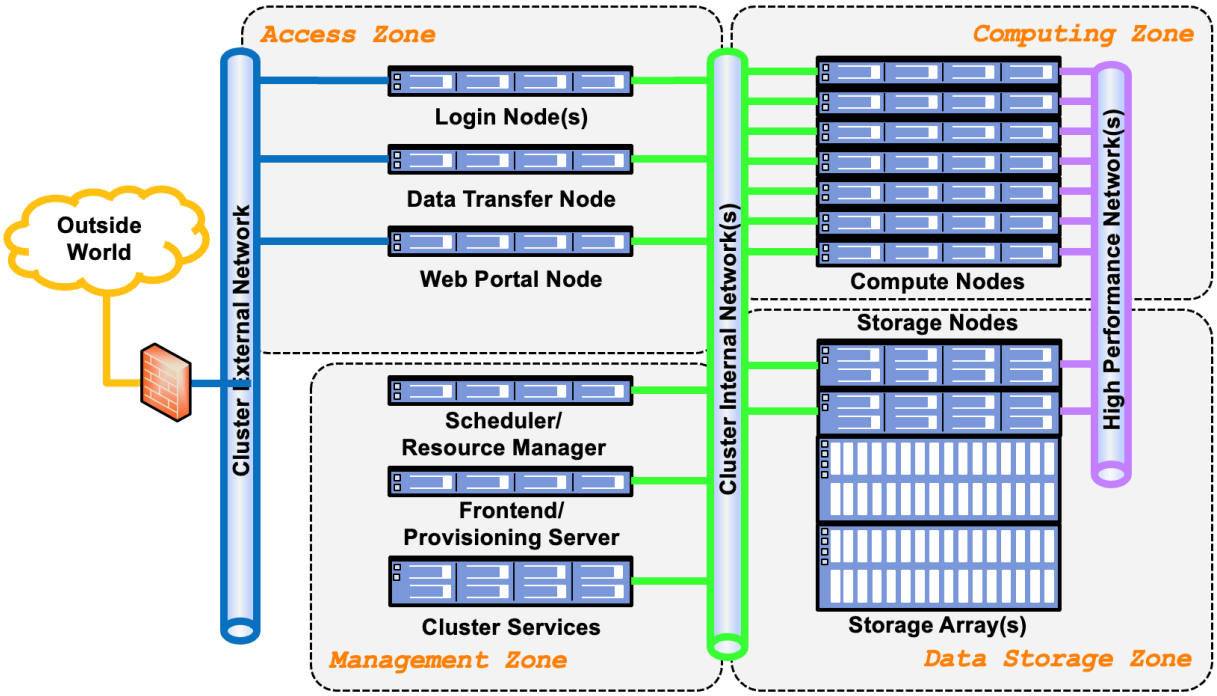

Fig. 1. HPC reference architecture

While HPC systems and AI data centers share many similarities, an AI data center is specifically designed for the training and inference of AI models and AI-powered applications. Co-design is commonly used in AI data centers to push the boundaries of system performance. For instance, AI training and inference accelerators are densely deployed alongside comprehensive ecosystems of training datasets, models, and software stacks that are built to enable rapid development and deployment. Additionally, data storage systems are optimized to support the data-intensive requirements of model training and inference, among other features.

NIST has conducted analyses of HPC threats and security postures (NIST SP 800-223 [3]) and developed the HPC security overlay (NIST SP 800-234 [4]) in collaboration with other government agencies, national laboratories, industry, and academic institutions. The HPC security overlay has been well-received by the HPC community. We intend to leverage NIST SP 800-223 and NIST SP 800-234 when developing security guidelines for an AI data center. To achieve that, this NIST Special Publication aims to conduct a threat and security gap analysis of AI data centers relative to HPC systems and to provide basic security recommendations. It shall lay the foundation for the future development of security guidelines for AI data centers.

For the rest of this document, we first identify key differences between AI data centers and HPC systems across system architecture, hardware, software stacks, workflows, and data storage systems. We then outline the major threats facing AI data centers. AI data center security challenges and potential solutions are included at last.

2. Reference Architecture and Differences

The following subsections introduce a reference architecture for AI data centers and high-level descriptions of individual zones. The subsequent gap analysis highlights the key differences between an AI data center and an HPC system.

2.1. AI Data Center Reference Architecture

An AI data center is specifically optimized for the training and inference of AI models. This optimization is evident in its hardware infrastructure, software stacks, workflows, and data storage systems. The AI data center reference architecture is illustrated in Fig. 2.

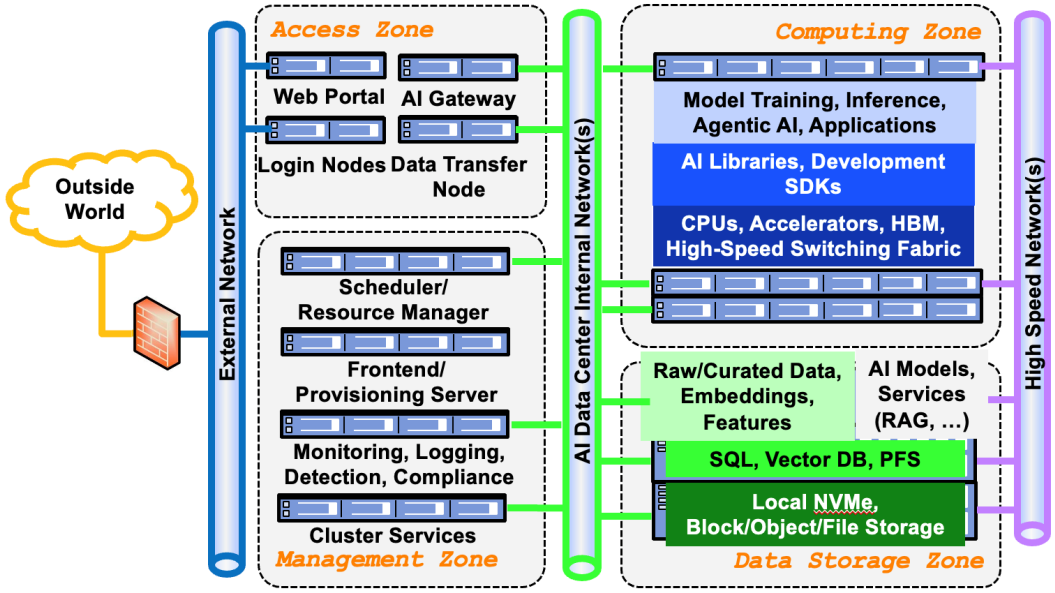

Fig. 2. AI data center reference architecture

An AI data center is structured into four functional zones:

1. **Computing Zone:** The Computing Zone provides key services for preprocessing the training dataset, training AI models, and providing inference and other AI services. Compute nodes in an AI data center are often equipped with accelerators to accelerate AI model training and inference (see Sec. 2.2.2). These are often used in conjunction with High-Bandwidth Memory (HBM) [5] for local data caching, key-value (KV) caching [6], model parameter caching, and related operations. High-speed networks are used for both scale-up and scale-out to build a unified AI supercomputing system. For instance, NVLink [7] can connect multiple accelerators within the same rack to scale up computing, while an InfiniBand-type [8] high-speed network connects computing nodes across different racks to scale out computing resources. This software stack (e.g., AI libraries, AI software development kits [SDKs], tools) enables efficient AI model training and inference.

2. **Data Storage Zone:** The Data Storage Zone provides a pivotal service for storing and accessing data since AI workloads consume and generate tremendous amounts of data. Like an HPC system, an AI data center storage system is multi-tiered to meet the extreme performance, capacity, and latency requirements of modern AI training and inference.

   Data storage systems typically consist of local disks, block storage, object storage, and/or file storage. Databases and high-performance PFSs run on top of them to provide fast data retrieval services. In addition to structured data, AI model training also consumes large amounts of unstructured data, such as video, audio, and PDF files. Cost-effective object-based storage is often adopted to store vast amounts of raw data in its native format.

   An AI data center storage system often acquires data from external storage systems for training and inference purposes and may share AI models and tokens with external parties. For example, retrieval-augmented generation (RAG) is an AI inference framework that improves the accuracy of large language models (LLMs) by retrieving data from external sources (e.g., company documents, live databases) in real time before generating a response. In the context of agentic AI, data storage systems may also retrieve and share data with external systems.

3. **Access Zone:** The Access Zone in an AI data center has significantly greater responsibilities than in an HPC system. In addition to the services provided by an HPC system, the Access Zone in an AI data center hosts an AI gateway to handle incoming AI inference and application requests. The AI gateway routes those requests to the appropriate servers, which promptly generate responses. Monitoring and logging user requests and AI-generated responses are often required for security and auditing purposes. The Access Zone provides an interface for data storage systems to retrieve external datasets and/or share internal datasets. All of these functions expose the Access Zone to greater threats and risks.

4. **Management Zone:** The Management Zone in an AI data center operates similarly to that of an HPC cluster. However, the Management Zone in an AI data center has greater responsibility for monitoring, logging, anomaly detection, and compliance checks. This is due to more frequent and diverse user access, increased data acquisition and sharing, more capable AI models, and heightened regulatory and compliance requirements for data and AI models. As a result, extensive monitoring, logging, and compliance checks become necessary. These functions are typically managed within the Security Operations Center (SOC), which is essential for protecting high-value assets, monitoring unique AI-related attack surfaces, and ensuring regulatory compliance. Overall, the Management Zone plays a critical role in maintaining the security posture of an AI data center.

While the composition and services offered by these zones resemble those found in an HPC system, AI data centers are specifically designed to support the unique requirements of AI workloads.

2.2. Computing Zone Hardware Infrastructure Difference

The hardware components in an AI data center computing zone differ from those in an HPC computing zone in several key areas.

2.2.1. Accelerator-Dense Infrastructure

An AI data center is characterized by a high concentration of specialized AI accelerators per server, per rack, and across the cluster. In these environments, multiple accelerators on the same compute node are often interconnected via high-bandwidth fabrics (e.g., NVLink [7]), while compute nodes are connected via high-speed networks (e.g., Spectrum-X [9], InfiniBand [8], Slingshot [10]). These high-speed connections enable the tight coupling of accelerators within each node and are scaled across racks to support large-scale model training and high-throughput, low-latency inference services.

2.2.2. Diverse AI Accelerators

AI accelerators are specialized hardware designed to accelerate AI training and inferencing. There are many types of AI accelerators, such as graphics processing units (GPUs) [11], Google's Tensor Processing Units (TPUs) [12], Neural Processing Units (NPUs) [13], AWS Trainium [14], and Language Processing Units (LPUs) [15]. Different accelerators have different design goals and use different technologies. Inference providers and model developers are increasingly developing custom silicon. For instance, TPUs are Google-developed application-specific integrated circuits (ASICs) that are specifically optimized for massive tensor operations in machine learning to offer high-efficiency and speed for TensorFlow. LPUs are designed for fast inference, while NPUs are designed for low-power, high-efficiency inference on edge devices.

2.2.3. Reduced-Precision Arithmetic

In AI data centers, floating-point representations in accelerators have shifted from high-precision (e.g., FP32 and FP64 for scientific computing) to lower-precision formats (e.g., FP16, BF16, and increasingly FP8) to maximize training speed, reduce memory bandwidth bottlenecks, and decrease power consumption. These formats use the IEEE 754 standard [16], which encodes numbers using a sign bit, an exponent (scale), and a significand/mantissa (precision) to represent a wide range of values for neural network weights and activations.

2.2.4. Use of High-Bandwidth Memory

During AI model training, the training data, model weights, and gradients are cached in memory that accelerators frequently access at high speed. In LLM inference, the requirements for memory size and bandwidth are even more demanding than training. In addition to requiring substantial memory to store KV pairs, the model parameters and KV caches must often be loaded onto accelerators to produce individual output tokens. This process necessitates a significant amount of memory bandwidth. High-bandwidth memory (HBM) meets this

requirement by leveraging advanced stacked memory technology that is tightly integrated with AI accelerators to deliver extremely high data throughput and low latency. Unlike traditional double data rate (DDR) memory attached to CPUs, HBM is co-packaged with the AI accelerator, enabling massive parallel memory channels and bandwidth critical for large-scale model training.

2.2.5. AI-Optimized High-Speed Networking

High-speed networking in AI data centers is a foundational component that enables large-scale, accelerator-dense infrastructure to function as a unified system rather than as isolated compute nodes. Modern AI workloads generate massive within-node (i.e., accelerator-to-accelerator) and cross-node data exchange for gradient synchronization, parameter updates, and checkpointing, which drives the adoption of high-speed interconnects with remote direct memory access (RDMA) [17] capabilities.

In AI data centers, large-scale AI model training requires frequent all-reduce operations and gradient exchanges across accelerators, creating east-west traffic with microbursts and congestion at scale. As a result, AI data center networking emphasizes extremely high bandwidth, RDMA efficiency, congestion control tuning, and scalability across tens of thousands of accelerators in addition to the short latency. Cost-effectiveness, performance isolation in a multi-tenant environment, and seamless integration with existing data center infrastructure are some factors to consider when choosing the right networking technology for an AI data center. For example, Ethernet is a scalable, cost-effective networking standard. RDMA over Converged Ethernet (RoCE) is a high-performance network protocol that enables RDMA directly over Ethernet networks.

2.3. Computing Zone Software Stack Difference

An AI data center has a vast ecosystem of libraries, tools, and development SDKs that facilitate model training and inference by leveraging the AI computing infrastructure.

2.3.1. System Software and Deep Learning SDK

These drivers facilitate model training and inference by utilizing the underlying accelerators and networking to achieve high-performance parallel computing. AI data center system software includes various tools, such as Compute Unified Device Architecture (CUDA) [18] for Nvidia GPUs, Radeon Open Compute (ROCm) [19] for AMD GPUs, and OpenVINO [20] for Intel accelerators. Cross-platform standards are also supported, such as OpenCL (Open Computing Language) [21] and SYCL [22].

2.3.2. Deep Learning Libraries

Deep learning libraries are software frameworks that facilitate and accelerate the training and validation of AI models. For instance, PyTorch [23] and TensorFlow [24] provide prebuilt

components (e.g., neural network layers, optimizers, GPU acceleration) to efficiently design, train, and validate complex AI models. Megatron Core [25] is a PyTorch-based library for composable training of large-scale generative AI. It provides GPU-optimized building blocks for training and post-training workflows, while DeepSpeed [26] provides a flexible optimization library that maximizes throughput and reduces memory usage. NVIDIA NeMo [27] provides an open suite of libraries with skills for accelerating AI agent specialization, optimization, and governance.

2.3.3. Resource Management and Orchestration

The workflows and processes in an AI data center are complex. The main challenge is to effectively coordinate accelerator-heavy workloads, high-speed networking, and distributed storage. While SLURM [28] is widely used in HPC-style AI clusters, Kubernetes [29] has been increasingly adopted for container runtime management in virtualized environments.

2.4. Data Storage Zone Difference

A data storage system in an AI data center is designed to support the end-to-end AI workflow, including raw data storage, data preprocessing, model training, and inference. AI training and inference tasks require significant data movement between memory, CPUs, accelerators, and data storage systems. The quality of AI models heavily depends on the volume and quality of their training data, which can be costly to collect, generate, and maintain. Therefore, a data storage system must offer extreme capacity, high performance, ample aggregated I/O bandwidth, and low latency, all while ensuring security and economic feasibility.

2.4.1. Use of Object Storage as the Backbone

Object storage manages information as flexible units called objects rather than in rigid file hierarchies or blocks. Each object bundles data, customizable metadata (i.e., contextual information), and a unique ID. Object storage offers substantially better scalability, resilience, and durability and can deliver large bandwidth to and from compute nodes. It achieves excellent performance by abandoning the notion of files and directories and only supports two fundamental operations: PUT and GET. It is ideal for storing unstructured data, such as videos, images, and backups. Object storage can handle petabyte- and exabyte-scale computing and support parallel reads with high aggregated bandwidth at relatively low cost.

2.4.2. Layered Data Storage System

Local NVMe [30], block, and file storage also support hot, warm, and cold data access to achieve performance, scalability, and cost-effectiveness. Various data services are built on physical data storage to support a variety of AI data formats and workloads, such as SQL engines for structured access, vector databases for semantic retrieval, PFS for high-performance workloads, and metadata services for governance.

Data exists in various formats and requires different data services during AI training, inference, and applications. For instance, AI workflows start with preprocessing raw data and converting it into tokens and features that are suitable for training AI models. AI model training requires large-batch access and frequent checkpointing. In contrast, AI inference handles frequent concurrent requests that generate many small reads. As another example, RAG [31] improves LLM accuracy and reduces hallucinations (i.e., false information) by pulling data from more up-to-date and domain-specific sources in response generation. RAG databases usually sit on an enterprise IT system and/or in the cloud. The data is tokenized, embeddings are created, and the embeddings are stored in a vector database. This process can take place in multiple ways, including via application programming interface (API) calls to a larger AI data center infrastructure or via some local infrastructure specific to the task.

2.5. Access Zone Difference

In an AI data center that provides AI services, an AI gateway serves as a centralized control point to provide a unified interface between external applications and AI services. In collaboration with the Management Zone, the AI gateway enhances the functionality of standard API gateways by incorporating features (e.g., intelligent AI workload routing, multimodal data integration, load balancing, token consumption tracking, rate limiting) and guardrail functionalities (e.g., content and sensitive information filtering, contextual grounding checks, reasoning checks).

Additionally, a data transfer node in an AI data center is crucial for managing high-speed data movement between external storage systems and internal high-performance storage. These nodes serve as secure, optimized gateways that facilitate rapid data ingestion and egress without interrupting AI model training or services. Since AI model training and services often rely on real-time data streaming and data acquisition from external sources, the security monitoring and logging features at the data transfer node are essential. For some high-security AI data centers, ingress and egress gateways can be rate-limited to further enhance the security posture.

2.6. Management Zone Difference

Monitoring, logging, security detection, and compliance functions in an AI data center are enhanced to protect against both traditional cybersecurity threats and AI-specific risks. These functions are placed across the entire AI infrastructure stack and are often part of the SOC function. They ensure that the massive datasets used to train high-value AI models will not be accessed by unauthorized entities. Prompts are monitored to detect malicious inputs that could hijack LLM behavior during inference. The security policy is strictly enforced, and sufficient monitoring and logging are in place to satisfy regulatory and compliance requirements.

2.7. Open Models and Datasets

Open models provide public access to components such as weights, parameters, code, and training data, allowing users to study, modify, and run them freely. Unlike closed or proprietary models, these open models can potentially promote transparency and various forms of collaborations, in addition to the ability to host models locally.

Open models can, in some instances, enable users to leverage advanced AI models without training them from scratch, significantly speeding up development and reducing costs. Users can easily access state-of-the-art pre-trained models, such as Transformers (e.g., LLaMA [32], Gemma [33]), for various tasks (e.g., text classification, named entity recognition, text generation). These open models can be fine-tuned on specific datasets for quick customization. Additionally, open models allow users to run the software locally, enhancing privacy and security.

3. AI Data Center Security Threats and Challenges

While many of the threats to HPC systems remain relevant, AI data centers introduce novel threats. The following subsections describe key security threats and challenges that will impact the design of the AI data center security guidelines.

3.1. AI Data Center Scalability Challenge

The rapid evolution of AI infrastructure means that newly built AI data centers can operate at a scale that dwarfs the world's most powerful traditional supercomputers. While the TOP500 [34] list has been the gold standard for HPC, modern AI data centers are now being engineered with accelerator density and interconnect bandwidth that make even the fastest exascale HPC systems appear modest. This divergence is driven by the ever-increasing demand for training large-scale AI models and providing AI services to the public. As this gap widens, managing an expansive, unified fabric of compute, memory, and high-speed storage becomes increasingly complex.

This unprecedented density creates a volatile operational environment in which hardware failures occur daily, sometimes hourly. Ensuring the continuity of massively distributed training jobs and real-time inference services becomes a monumental task. A single node failure can trigger a cascade that halts a multi-week training run. Consequently, building resilient infrastructure is no longer just about hardware durability but also about developing sophisticated software orchestration and checkpointing mechanisms that can bypass faulty hardware without compromising the security or integrity of the workload. This complexity and speed also pose tremendous challenges for monitoring, logging, and real-time anomaly detection.

3.2. Data Quality, Provenance, and Security Challenges

Data quality significantly affects both the speed of training and the quality of trained AI models. Collecting high-quality data, tracking its provenance and usage history, and preserving it for the long term are major investments in the AI development life cycle. Designing an AI data center that effectively meets these requirements is a challenge.

In addition, datasets and trained AI models are highly valuable. Data exfiltration, model theft, and data poisoning pose major threats to AI data centers. Categorizing the security requirements for raw data, intermediate data, and AI models (e.g., model weights) and protecting them effectively is of great importance.

3.3. Regulatory and Compliance Challenges and Logging, Monitoring, and Audit-Ready Challenges

The governance of datasets used for AI model training and inference has become a primary hurdle for infrastructure providers since modern workloads frequently ingest massive amounts of sensitive data that is often subject to global and local regulations. For example, the privacy

controls of the General Data Protection Regulation (GDPR) [35] and the California Consumer Privacy Act (CCPA) [36], the patient confidentiality requirements of the Health Insurance Portability and Accountability Act (HIPAA) [37], and the government security protocols for Controlled Unclassified Information (CUI) [38] require rigorous data provenance and isolation. In an AI context, this is particularly difficult because the black-box nature of AI models can make it difficult to prove that sensitive information has not been memorized or leaked during training or inference. Consequently, data centers must implement multi-layered security architectures that include end-to-end encryption, strict identity and access management (IAM), input and output token monitoring, and sometimes physical air-gapping to ensure that training sets remain untainted and legally compliant.

Building an AI data center that meets these aims represents significant engineering and administrative challenges. Traditional auditing against system or center requirements relies on point-in-time snapshots, but the dynamic nature of AI (e.g., continuous learning, frequent model updates, distributed storage) requires automated, real-time logging at every data touchpoint. Operators need to maintain chain-of-custody records for data from initial ingestion to its ultimate influence on a model's weights. This requires a sophisticated orchestration layer that is capable of generating granular reports on data residency, processing logs, and administrative access without degrading the massive computational throughput required for AI. Ultimately, being able to meet these demands is critical to the data center's design and operations.

3.4. Virtual Environment Threats

To accommodate complex AI workflows, an AI data center often abstracts physical, high-performance hardware (e.g., GPUs, TPUs, specialized networking) into a flexible, on-demand resource pool. These environments allow organizations to run demanding AI workloads (e.g., LLM training) using cloud-native technologies (e.g., containers, Kubernetes) on premises or in specialized, high-performance cloud environments. This flexibility enables dynamic scaling to meet peak demands without the overhead of manual hardware provisioning. Virtual environments are also widely used in agentic AI to provide an isolated workspace (i.e., sandboxes) for testing and secure code execution by providing an isolated workspace.

However, this layer of abstraction introduces a complex surface for security vulnerabilities. Sharing hardware across multi-tenant workloads must account for the risk of breakout attacks in which a malicious actor escapes a container or virtual machine to access the underlying host or adjacent memory. Various technologies (e.g., virtualized networking stacks, orchestration APIs) could also be exploited to intercept sensitive training data or manipulate model weights, and side-channel attacks and hypervisor vulnerabilities could compromise the integrity of the AI workflow.

3.5. Denial-of-Service (DoS) Threats

Adversaries may target machine learning systems by intentionally crafting inputs that require substantial computation, thereby flooding the system with requests that can degrade or shut

down service during inference. These denial-of-service (DoS) attacks can also target streaming data during both training and inference. Rather than relying solely on static, pre-collected datasets, AI training on streaming data involves continuously updating machine learning models as data is generated. This approach allows models to adapt in real time to changing patterns, which is critical for time-sensitive applications, such as financial fraud detection and recommendation engines. The attacks can target streaming data and use a DoS attack to prevent the training process from receiving it in time. Finally, attacks against the data center's operational technology (OT) (e.g., power and cooling systems) can also lead to denials of service. AI data center OT security is covered in a separate ongoing effort.

3.6. Prompt-Based Exploitation Threats

Concerns regarding AI security — especially those related to prompt-based exploitation — cannot be addressed solely through the models themselves. A robust infrastructure is essential to effectively mitigate these risks. Organizations can reduce the likelihood of prompt injection attacks by implementing technical measures, such as context-based access controls (CBAC) and input/output guardrail pipelines in the data center.

The following attacks can exploit AI input prompts:

- **Model distillation attack:** A model distillation attack (or model extraction attack) is a security threat in which an attacker queries an AI model to generate a dataset of input-output pairs. This dataset is then used to train a smaller, cheaper model that mimics the original's behavior. This intellectual property theft is often carried out through large-scale, automated queries to clone specialized AI models, as seen in cases flagged by Anthropic [39]. AI data centers that host an AI model must strive to detect and mitigate such attacks.
- **AI model ontology discovery attack:** An AI model ontology discovery attack is a reconnaissance technique with which adversaries map a target AI model's internal structure (e.g., output classes, object types, capabilities) to enable targeted, downstream attacks. By repeatedly querying the AI model, attackers may force it to reveal its classification system, enabling them to craft an effective targeted attack.
- **AI black-box attack:** In an AI black-box attack, attackers manipulate an AI model's input without knowing its internal structure or parameters. Attackers use repeated queries to observe changes in output and reverse-engineer vulnerabilities to bypass security controls or cause misclassifications.
- **Prompt injection attack:** In a prompt injection attack, attackers feed malicious inputs to bypass the AI's security guardrails or override original instructions. By manipulating input prompts, attackers trick models into performing unintended actions, disclosing private information, or behaving inappropriately. LLMs are particularly vulnerable since they are typically required to handle untrusted input, and applying adequate access control is difficult.

3.7. Supply Chain Threats

- **Hardware supply chain threat:** AI data center components are often sourced globally, which increases exposure to supply chain risks. Malicious actors, including nation-states, target physical components, servers, accelerators, and networking gear to install backdoors, steal intellectual property, or enable sabotage. The intense demand for AI hardware has compressed innovation cycles, creating a high-volume, concentrated, and poorly secured supply chain.
- **Software supply chain threat:** An AI data center runs a complex software ecosystem that mixes proprietary and open-source software suites whose rapid development and adoption have introduced supply chain risks. For instance, in December 2022, the PyTorch open-source machine learning framework was targeted by a supply chain attack involving a malicious dependency named torchtriton [40]. In 2024, the cybercriminal group NullBulge targeted AI-driven applications and gaming communities by poisoning legitimate code repositories on platforms such as Hugging Face [41] and GitHub [42].
- **AI dataset and model supply chain threat:** AI dataset and model supply chain attacks are sophisticated cyber attacks in which malicious actors poison, corrupt, or manipulate the training data and pre-trained models used to build AI systems. The scarcity of training datasets and the high cost of model training force developers to rely on third-party pre-trained models and datasets, which creates an expansive, under-secured attack surface. For instance, incidents involving malicious models on platforms such as Hugging Face have increased significantly [43]. A poisoning data attack can corrupt or alter data, skew the training process, and lead to unreliable models.

3.8. Multi-Tenant Threats and the Three-Way "Trust Dilemma"

AI data centers are designed to train and provide AI services for multiple tenants. AI workloads are scheduled intelligently, and accelerators may be shared in fractional units to maximize utilization. The shared infrastructure (e.g., compute, storage, and network) across both on-premises and cloud-based AI data centers introduces risks ranging from data breaches to physical infrastructure tampering. Cross-tenant data leakage, side-channel attacks, noisy-neighbor resource contention, and malicious power/thermal attacks may occur. As AI demands intensify, these shared environments will be increasingly targeted by sophisticated, AI-driven attacks.

Deploying proprietary models on a multi-tenant infrastructure introduces a complex, three-way "Trust Dilemma" among tenants/users, model owners, and AI data center operators [44]. Tenants/users need to ensure that their applications and data remain private and secure. Model owners need to protect their model weights and algorithms from sabotage. Infrastructure operators must protect AI data centers from potential malicious code and tenant-driven attacks. Establishing trust and providing security in a multi-party environment is challenging.

3.9. Insider Threats

Insider threats in AI data centers pose a critical risk. Trusted employees, contractors, or administrators could abuse their access to steal, sabotage, or unintentionally expose sensitive AI models and training data. The recent advancement in agentic AI and the autonomous nature of the agents make them “insiders” as well.

AI models are high-value intellectual property. The data volumes are massive, and the speed at which data can be exfiltrated is fast. AI hardware, software, and tools are often adopted faster than security protocols can be developed. Model theft, data poisoning, unintentional data leakage, and the potential sabotage of ML systems are among the threats that insiders can pose.

3.10. Silent Data Corruption (SDC) Threat

Silent data corruption (SDC) [45] occurs when a CPU or an accelerator/GPU produces an incorrect result from an arithmetic operation, leading to data loss or corruption. Many defective chips are escaping today’s manufacturing tests. SDCs caused by test escapes can lead to incorrect floating-point calculations, invalid inferences, flawed training data, or skewed results in AI applications.

SDC can be mitigated using software-level defenses, proactive hardware screening, and/or data redundancy. For instance, Google uses a multi-layered, proactive approach that combines software-level checks, rigorous hardware screening, and architectural redundancy [45].

3.11. Expanding and Evolving Attack Surface Threats and AI-Powered Attacks

With the introduction of new hardware, software, AI models, and agents, as well as the complex interactions among those components, the attack surface continues to expand and evolve. Emerging AI-powered attacks further broaden the attack surface and expose vulnerabilities. The measurement of time-to-exploit (TTE) — which is the time gap between the public disclosure of common vulnerabilities and exposures (CVEs) and the first confirmed in-the-wild exploitation — shows that the mean TTE has decreased by more than 1,000 since 2018 [46]. AI lowers the barrier to entry for attackers, enables the rapid development and deployment of zero-day attacks, and makes attacks more cost-effective. Agentic AI-based attacks further enable fully autonomous attacks. Developing effective security tools that keep pace with change is a major security challenge.

3.12. Multiple Data Center Training Threat

Multi-data-center training overcomes the physical, power, and cooling constraints of a single data center. It enables the fast training of increasingly massive models, provides fault tolerance, and achieves efficient resource utilization. For example, Microsoft builds multiple data centers at different locations and connects them using an AI Wide Area Network (AI WAN) of dedicated

fiber-optic cables [47]. This design allows data to travel at nearly the speed of light and congestion-free, enabling multiple sites to cooperate on near real-time model training.

Multiple data centers introduce unique technical, security, and physical threats due to their reliance on high-speed, long-distance networking and synchronized operations. Connecting multiple data centers requires exposing internal networks to public infrastructure. Attackers can target these connections, intercept traffic, or launch man-in-the-middle attacks.

3.13. Firmware Integrity Threat

Firmware integrity threats are among the most significant threats to AI data center security. Components operate below the operating system and can bypass many traditional security controls, such as Baseboard Management Controller (BMC) (see Sec. 3.13.1), Unified Extensible Firmware Interface (UEFI), Basic Input/Output System (BIOS), Network Interface Card (NIC), GPUs, and storage controllers. They run with high privileges, persist across OS reinstallations, and even survive disk replacement or software recovery. The following subsections review some of the firmware security threats that can be significant in an AI data center.

3.13.1. Baseboard Management Controller (BMC) Threats

A BMC is a specialized service processor embedded on the motherboard of compute and data storage nodes that enables administrators to remotely monitor, power cycle, and troubleshoot hardware. In an AI data center, BMCs are connected to the Management Zone via out-of-band management networks and operate independently of the host OS. BMCs can modify BIOS settings, apply firmware updates, and control boot behavior. They also collect detailed telemetry, like temperature, power draw, and fan speed. This high-level access introduces potential security risks.

Multiple vulnerabilities and security risks have been identified in [48], such as credential leakage via side channels, full remote access via insecure APIs, memory exploits, and pivoting to host systems.

3.13.2. Unified Extensible Firmware Interface (UEFI) Threat

A UEFI attack targets the low-level software that initializes hardware and boots an operating system. Because UEFI operates at the highest privilege level on a motherboard, these attacks can bypass operating system security and survive OS wipes, hard drive replacements, and factory resets.

UEFI attacks can reside on a hard drive's EFI system partition or implant code in the motherboard's serial peripheral interface (SPI) flash memory. UEFI attacks become feasible in an AI data center when attackers gain full access to a server via bare metal as a service or a successful BMC attack.

3.14. AI Accelerator Vulnerability

The emphasis on high-throughput and extensive parallel computation in AI data centers has expanded the attack surface across AI accelerators (e.g., GPU hardware, firmware, associated drivers). This complex ecosystem often lacks sufficient monitoring and can introduce unique vulnerabilities. For example, inadequately clearing video random-access memory (VRAM) after a task completes may allow subsequent tenants to access remnants of sensitive data on shared hardware. Furthermore, emerging research [49] demonstrates that Rowhammer-style memory exploits can effectively target GPUs to manipulate machine learning models and lead to substantial degradation in model accuracy.

4. AI Data Center Security Posture and Possible Solutions

In this section, we identify security measures that help to mitigate the threats outlined in Section 3. Although this is not a comprehensive list of potential strategies, it provides a foundation for enhancing AI data center security.

4.1. Strengthening the Access Zone Protection

The Access Zone serves as the gateway between an AI data center and the external world, providing a crucial interface for both users and system administrators. Given that data centers host highly valuable and sensitive AI models and data, they face sophisticated threats from malicious attackers, industrial espionage, and state-sponsored actors.

Although AI gateways and traditional API gateways share many similarities, they also exhibit significant differences that lead to distinct security vulnerabilities [50]. For example, authentication issues, API inventory management, and security configuration remain primary vulnerabilities in the AI gateway. Other vulnerabilities (e.g., Broken Object Level Authorization, unrestricted resource consumption) differ from those found in traditional APIs and can be exacerbated by the capabilities of AI models. For instance, the input prompts for AI models tend to be longer and more general. Traditional security mechanisms (e.g., object-level and function-level authorization) are often inadequate for AI APIs.

Attackers aiming to exploit AI models may issue legitimate queries or API calls to the AI gateway, potentially executing a distillation attack. To mitigate these threats, AI-powered mechanisms are essential, such as guardrail AI prompt pre-evaluation. A key initial step is to log users' queries and API calls along with the corresponding responses from AI models to effectively detect suspicious activity. Additionally, the Access Zone must be properly configured and equipped with sufficient resources to manage high volumes of requests and API calls.

Of note, an AI gateway may also operate within the Computing Zone, depending on the specific application scenarios (e.g., to provide API services for agentic AI applications). Regardless of its location, the AI gateway serves as a critical security chokepoint and should be rigorously monitored and secured.

4.2. Providing Strong Monitoring, Logging, Security, and Compliance Oversight Capabilities

Proactive oversight is essential due to the high-stakes nature of AI workloads — including the potential for data and model exfiltration — and the significant financial and reputational damage of breaches. Data and model access and versioning history can be used for provenance and anomaly detection.

Compliance and audit requirements are common in an AI data center. Regulations in some instances can mandate rigorous data privacy, model transparency, and auditability, and non-compliance can lead to penalties. Maintaining audit readiness provides the necessary evidence and audit trail to demonstrate compliance, which is essential for mitigating ethical, legal, and operational risks and enhancing users' trust.

An AI data center is a large-scale, complex AI computing infrastructure shared by many users/tenants. Monitoring their usage and access patterns, data and model access patterns, and lateral inter-process communication patterns helps detect anomalies and potential attacks.

4.3. Zero-Trust vs. Nurturing Trust

Traditional network security measures (e.g., firewalls, virtual private networks) are often inadequate for protecting against internal threats, compromised credentials, or supply chain attacks. In contrast, zero trust is a security framework that assumes that breaches are inevitable and eliminates the assumption of implicit trust. Instead, it requires strict and continuous authentication and authorization for every user, device, application, and workflow. Every request is treated as a potential threat, making it crucial for securing sensitive training data and proprietary AI models from unauthorized access, lateral movement, and data leakage. Adopting a zero trust approach prevents an attacker from accessing or stealing the AI model itself, even if one component is compromised.

Meanwhile, hardware security has seen significant advancements, particularly in hardware root of trust (RoT) and confidential computing. A hardware RoT is an immutable security component (e.g., Trusted Platform Module, secure enclave) that is inherently trusted to perform cryptographic functions, ensure device integrity, and secure boot processes. It serves as the foundation for all system security by verifying software before it runs, preventing tampering, protecting cryptographic keys, and establishing a chain of trust. It is effective against hardware and firmware attacks, such as UEFI attacks that target the low-level firmware that boots compute nodes.

Confidential computing is another security technology that brings trust to the computing platform. It protects data in use by processing it within a hardware-based trusted execution environment. This ensures that data remains encrypted and isolated from unauthorized access, including that of AI data center operators, operating systems, and hackers.

By leveraging the latest advancements in hardware security and implementing a zero trust principle in their design and operations, organizations can strike the right balance between enhanced security and privacy and high data center performance.

4.4. Enabling Human-In-The-Loop Oversight and Governance Capability

In an AI data center, certain activities can have serious consequences, such as the unintended deployment of incorrect or unapproved models to production. Human-in-the-loop (HITL) governance can mitigate these risks. Alongside traditional user authentication and authorization, the AI data center should incorporate a technological mechanism that enables human oversight and governance and requires final approval for critical actions. For instance, human oversight should ensure that permissions, version control, and proper approvals are in place before models are promoted to production. Insufficient controls within the machine learning and systems development life cycle can lead to the accidental deployment of incorrect or unapproved models.

4.5. Developing a Robust and Fault-Tolerant Workflow

A fault-tolerant workflow helps an AI data center manage complex hardware and software glitches, ensure the continuous operation of intensive AI training and inference jobs, and mitigate financial losses from work interruptions. Accelerator failures, disk failures, and software glitches are common in large-scale AI data centers, and large-scale model training can take weeks or months. AI jobs must be able to continue amid these disruptions to shorten training duration and minimize total cost.

A fault-tolerant workflow can be achieved through a multi-layered approach, including hardware redundancy, active workflow monitoring, and intelligent checkpointing and recovery. Disk failures can be hidden by the data storage system, whereas a GPU failure or a hung process can trigger local recovery to restart the failed tasks without affecting overall model training. For example, the dynamic replay technique recomputes only the portions of a training step that were lost or corrupted when a fault occurred rather than restarting from a checkpoint. If a node fails mid-step, surviving nodes can replay only the affected micro-batches or layers. During forward and backward passes, intermediate tensors (e.g., activations, gradients) can be cached or recomputed on demand. These techniques allow for faster recovery with minimal impact on job progress.

4.6. Security-by-Design and Verification-by-Design Approach to AI Data Center

A security-by-design approach embeds security into the AI data center life cycle from the beginning rather than adding it as an afterthought. This lowers remediation costs, reduces the risk of data breaches, enhances product quality, and eases regulatory compliance. It also builds customer trust and reduces long-term maintenance by minimizing vulnerabilities early.

A verification-by-design approach incorporates rigorous testing, monitoring, and compliance checks from the earliest stages of AI data center development. By providing continuous, real-time monitoring and compliance results, AI data centers can offer unprecedented transparency into security and compliance to foster trust among users, operators, and exporters.

5. Conclusions

Securing an AI data center is a complex task due to several factors. These include the facility's large size; high-performance requirements; the diverse and intricate hardware, software, and applications involved; varying security, regulatory, and compliance standards; the shared nature of resources; and the ongoing evolution of AI data centers. Currently, there is a lack of established standards, guidelines, and best practices specifically for AI data center security.

This publication aims to establish a foundation for standardizing and promoting the sharing of information and knowledge regarding AI data center security. It introduced a zone-based reference architecture for AI data centers and highlights common features across most AI data center systems. This publication also analyzed threats to AI data centers, considered security postures and associated challenges, and provided recommendations to improve security measures.

745 **References**

746

[1] "AI Action Plan," [Online]. Available: https://www.ai.gov/action-plan.

[2] J. Dongarra, D. Reed and D. Gannon, "Ride the Wave, Build the Future: Scientific Computing in an AI World," [Online]. Available: https://cloud4scieng.org/2026/02/06/ride-the-wave-build-the-future-scientific-computing-in-an-ai-world/.

[3] Y. Guo, R. Chandramouli, L. Wofford, R. Gregg, G. Key, A. Clark, C. Hinton, A. Prout, A. Reuther, R. Adamson, A. Warren, P. Bangalore, E. Deumens and C. Farkas, "NIST SP 800-223: High-Performance Computing Security: Architecture, Threat Analysis, and Security Posture," National Institute of Standards and Technology, 2024.

[4] Y. Guo, J. Licat, J. Neel, G. Key, J. Waterman, I. Lee, C. Hinton, D. Shrade, A. Prout, A. Reuther, T. Bohrer, K. Ishisoko, K. Earley, A. Warren, T. DeNardo, I. Czarnezki and E. Deumens, "NIST SP 800-234: High-Performance Computing (HPC) Security Overlay," National Institute of Standards & Technology, 2026.

[5] "High Bandwidth Memory," [Online]. Available: https://en.wikipedia.org/wiki/High_Bandwidth_Memory.

[6] "KV Caching Explained: Optimizing Transformer Inference Efficiency," [Online]. Available: https://huggingface.co/blog/not-lain/kv-caching.

[7] NVIDIA, "NVLink and NVLink Switch," [Online]. Available: https://www.nvidia.com/en-us/data-center/nvlink/.

[8] "InfiniBand," [Online]. Available: https://en.wikipedia.org/wiki/InfiniBand.

[9] NVIDIA, "NVIDIA Spectrum-X Ethernet Networking Platform," [Online]. Available: https://www.nvidia.com/en-us/networking/spectrumx/.

[10] "Slingshot," [Online]. Available: https://www.glennklockwood.com/garden/Slingshot.

[11] "What is a GPU?," [Online]. Available: https://aws.amazon.com/what-is/gpu/.

[12] "Tensor Processing Units (TPUs): Engineered for next-generation AI," [Online]. Available: https://cloud.google.com/tpu.

[13] "What is a neural processing unit (NPU)?," [Online]. Available: https://www.ibm.com/think/topics/neural-processing-unit.

[14] "AWS Trainium," [Online]. Available: https://aws.amazon.com/ai/machine-learning/trainium/.

[15] Groq, "What is a Language Processing Unit?," [Online]. Available: https://groq.com/blog/the-groq-lpu-explained.

[16] "IEEE 754," [Online]. Available: https://en.wikipedia.org/wiki/IEEE_754.

[17] "Remote direct memory access," [Online]. Available: https://en.wikipedia.org/wiki/Remote_direct_memory_access.

[18] "NVIDIA CUDA Toolkit," [Online]. Available: https://developer.nvidia.com/cuda/toolkit.

[19] "AMD ROCm Software," [Online]. Available: https://www.amd.com/en/products/software/rocm.html.

[20] "OpenVINO," [Online]. Available: https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html.

[21] "OpenCL for Parallel Programming of Heterogeneous Systems," [Online]. Available: https://www.khronos.org/opencl/.

[22] "SYCL," [Online]. Available: https://www.khronos.org/sycl/.

[23] "PyTorch," [Online]. Available: https://pytorch.org/.

[24] "TensorFlow," [Online]. Available: https://www.tensorflow.org/.

[25] "NVIDIA Megatron Core," [Online]. Available: https://developer.nvidia.com/megatron-core.

[26] "DeepSpeed," [Online]. Available: https://www.microsoft.com/en-us/research/project/deepspeed/.

[27] "NVIDIA NeMo," [Online]. Available: https://www.nvidia.com/en-us/ai-data-science/products/nemo/.

[28] SLURM, "SLURM Workload Manager," 2024. [Online]. Available: https://slurm.schedmd.com/documentation.html.

[29] Kubernetes, "Kubernetes," 2024. [Online]. Available: https://kubernetes.io/.

[30] "NVM Express," [Online]. Available: https://en.wikipedia.org/wiki/NVM_Express.

[31] "Retrieval-augmented generation," [Online]. Available: https://en.wikipedia.org/wiki/Retrieval-augmented_generation.

[32] "Llama: Build on your own terms," [Online]. Available: https://developer.meta.com/ai/.

[33] "Gemma," [Online]. Available: https://deepmind.google/models/gemma/.

[34] "TOP500," [Online]. Available: https://top500.org/.

[35] "General Data Protection Regulation," [Online]. Available: https://gdpr-info.eu/.

[36] "California Consumer Privacy Act (CCPA)," [Online]. Available: https://oag.ca.gov/privacy/ccpa.

[37] "Health Information Privacy," [Online]. Available: https://www.hhs.gov/hipaa/index.html.

[38] "Controlled Unclassified Information (CUI)," [Online]. Available: https://www.archives.gov/cui.

[39] "Detecting and preventing distillation attacks," [Online]. Available: https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks.

[40] P. Foundation, "Compromised PyTorch-nightly dependency chain between December 25th and December 30th, 2022," December 2022. [Online]. Available: https://pytorch.org/blog/compromised-nightly-dependency/.

[41] "The AI community building the future," [Online]. Available: https://huggingface.co/.

[42] J. Walter, "NullBulge: Threat Actor Masquerades as Hacktivist Group Rebelling Against AI," July 2024. [Online]. Available: https://www.sentinelone.com/labs/nullbulge-threat-actor-masquerades-as-hacktivist-group-rebelling-against-ai/.

[43] K. Zanki, "Malicious ML models discovered on Hugging Face platform," [Online]. Available: https://www.reversinglabs.com/blog/rl-identifies-malware-ml-model-hosted-on-hugging-face.

[44] "Confidential Computing for AI," [Online]. Available: https://docs.nvidia.com/ai-enterprise/planning-resource/ai-factory-white-paper/latest/confidential-computing-for-ai.html.

[45] Google, "Silent Data Corruption," [Online]. Available: https://support.google.com/cloud/answer/10759085?hl=en.

[46] "From Vulnerability to Exploitation," [Online]. Available: https://zerodayclock.com/.

[47] "From Wisconsin to Atlanta: Microsoft connects datacenters to build its first AI superfactory," [Online]. Available: https://news.microsoft.com/source/features/ai/from-wisconsin-to-atlanta-microsoft-connects-datacenters-to-build-its-first-ai-superfactory/.

[48] "Analyzing Baseboard Management Controllers to Secure Data Center Infrastructure," [Online]. Available: https://developer.nvidia.com/blog/analyzing-baseboard-management-controllers-to-secure-data-center-infrastructure/.

[49] C. Lin, J. Qu and G. Saileshwar, "GPUHammer: Rowhammer Attacks on GPU Memories are Practical," in *34th USENIX Security Symposium (USENIX Security 25)*, Seattle, WA, 2025.

[50] "OWASP API Security Top 10," [Online]. Available: https://owasp.org/API-Security/editions/2023/en/0x11-t10/.

747
748