Check for updates

NIST Interagency Report
NIST IR 8476

3rd High-Performance Computing Security Workshop

Joint NIST-NSF Workshop Report

Yang Guo
Jeremy Licata
Victoria Pillitteri
Sanjay Rekhi
Robert Beverly
Xin Yuan
Gary Key
Rickey Gregg
Stephen Bowman
Catherine Hinton
Albert Reuther
Ryan Adamson
Aron Warren
Purushotham Bangalore
Erik Deumens
Csilla Farkas

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8476

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

NIST Interagency Report
NIST IR 8476

3rd High-Performance Computing Security Workshop

Joint NIST-NSF Workshop Report

Yang Guo
Jeremy Licata
Victoria Pillitteri
Sanjay Rekhi
Computer Security Division
Information Technology Laboratory

Robert Beverly
National Science Foundation

Xin Yuan
Florida State University

Gary Key
Rickey Gregg
Stephen Bowman
HPCMP & NIWC Atlantic

Catherine Hinton
Los Alamos National Laboratory

Albert Reuther
MIT Lincoln Laboratory

Ryan Adamson
Oak Ridge National Laboratory

Aron Warren
Sandia National Laboratories

Purushotham Bangalore
University of Alabama

Erik Deumens
University of Florida

Csilla Farkas
University of South Carolina

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8476

September 2023

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

Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified 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.

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 2023-09-20

How to Cite this NIST Technical Series Publication:
Guo Y, Licata J, Pillitteri V, Rekhi S, Beverly R, Yuan X, Bowman S, Gregg R, Key G, Hinton C, Reuther A, Adamson R, Warren A, Bangalore P, Deumens E, Farkas C (2023) 3rd High-Performance Computing Security Workshop: Joint NIST-NSF Workshop Report. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) NIST IR 8476. https://doi.org/10.6028/NIST.IR.8476

Author ORCID iDs
Yang Guo: 0000-0002-3245-3069
Jeremy Licata: 0000-0001-8793-5471
Victoria Pillitteri: 0000-0002-7446-7506
Sanjay Rekhi: 0009-0008-8711-4030

Contact Information
ir8476-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

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

Abstract

High-performance computing (HPC) is a vital computational infrastructure for processing large data volumes, performing complex simulations, and conducting advanced machine learning model training. As such, HPC is a critical component of scientific discovery, innovation, and economic competitiveness. Cybersecurity thus plays an important role in HPC by safeguarding against abuse and misuse and ensuring data and research integrity. However, HPC systems often have unique hardware, software, and user environments that pose distinct cybersecurity challenges. This collaborative workshop gathered stakeholders from government, academia, and industry to discuss community needs, ongoing activities, and future directions in HPC security. This public workshop report provides detailed summaries of technical sessions, key takeaways from breakout sessions, and a summary of the keynote presentations.

Keywords

high-performance computing (HPC); HPC security; HPC security posture; security guidance; security compliance.

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.

Table of Contents

List of Tables

List of Figures

Preface

This report is intended to capture the activities and pertinent topics from the 3rd High-Performance Computing Security Workshop for public record and dissemination. The content of this report has been assimilated from individual workshop session scribe notes and edited for clarity and uniformity. As such, no claims are made as to the completeness or accuracy of the content herein. Rather, this report should be viewed as a summary of discussion-worthy points from the workshop. Further, this report does not reflect official National Institute of Standards and Technology (NIST), National Science Foundation (NSF), or federal positions or policies. All listed talks and participants provided consent for their names and session content to appear in this document.

1. Introduction: Workshop Objective, Participants, and Agenda

High-performance computing (HPC) serves as a fundamental computing infrastructure that drives scientific discovery, innovation, and economic competitiveness. The ability to process large volumes of data, carry out complex simulations, and conduct sophisticated machine learning (ML) model training at high speeds is essential for upholding the Nation's global competitiveness. Cybersecurity plays an important role in HPC by protecting against misuse, ensuring availability, and providing data integrity. Furthermore, the intellectual property embedded in algorithms, models, and HPC infrastructure design and architecture, as well as valuable and sensitive data necessitate comprehensive security measures.

HPC systems differ in the applications they run, the missions they support, and their system architectures and designs. The unique hardware, software, research, and users in the HPC ecosystem pose cybersecurity challenges that are frequently distinct from non-HPC environments. Each HPC system may possess unique security requirements and follow specific security guidance that tailored security solutions. For instance, HPC systems that belong to the U.S. Government must comply with the NIST Risk Management Framework (RMF), and authorizations and audits are required for system operation. In contrast, HPC systems that are supported by the National Science Foundation (NSF) rely on managing operators to secure the systems in a manner that is aligned with the science and discovery mission of the infrastructure. While different strategies are employed across the broader HPC ecosystem, security remains a crucial component to achieving the anticipated benefits of HPC.

The NIST HPC Security Working Group (WG) has been leading efforts to develop comprehensive and reliable security guidance for HPC systems. In pursuit of its mission and to engage the greater HPC scientific community, NIST collaborated with the NSF to host the 3rd High-Performance Computing Security Workshop on March 15 – 16, 2023, at the National Cybersecurity Center of Excellence (NCCoE) in Rockville, Maryland. The workshop gathered the community's needs and feedback, facilitated intellectual discourse, reported and reflected on ongoing activities in the HPC Security WG, and defined and discussed future directions with stakeholders from industry, academia, and government.

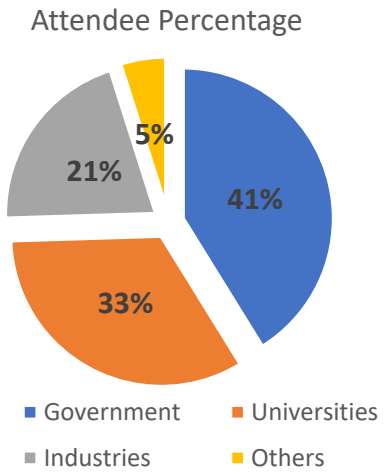

Fig. 1. Workshop attendee composition

The in-person two-day workshop attracted 102 attendees, with 41% representing government agencies, 33% from academia, 21% from industry, and 5% from other sectors (refer to **Fig. 1**). Among the attendees, over 50% held key roles in HPC security by serving as Chief Information Security Officers (CISOs), security engineers, security controls assessors (SCAs), HPC architects, and system engineers. The workshop encompassed three keynote talks, five technical sessions, multiple interactive panel discussions, and four breakout sessions with specific discussion focuses. Distinguished speakers were drawn from government agencies, national laboratories, NSF-funded open science supercomputing centers, universities, and industries. The workshop agenda and speakers' details can be found in **Table 1** and **Table 2**.

The rest of the workshop report is organized as follows:

- Section 2 covers highlights and summaries of individual sessions.
- Section 3 reports on the key takeaways of breakout sessions.
- Section 4 includes a summary of the keynotes.

Table 1. 3rd High-Performance Computing Security Workshop Agenda — Day 1

Start Time (ET)	End (ET)	Session/Talk
8:30 a.m.	9:00 a.m.	Welcome • Matthew Scholl, CSD Chief, NIST
9:00 a.m.	9:45 a.m.	Keynote 1: The NSF HPC Security Landscape: Research Challenges to Production Capabilities • Dr. Rob Beverly, NSF
9:45 a.m.	10:00 a.m.	Break
10:00 a.m.	11:45 a.m.	HPC Architecture and Security Posture • Chair: Rickey Gregg, HPCMP • Virtual Private HPC as a Software Development Sandbox ○ Andrew Prout and Albert Reuther, MIT LL • HPC Diskless Architectures ○ Aron Warren, SNL

Start Time (ET)	End (ET)	Session/Talk

		• The 'S' in HPC Stands for Security
		  o Ryan Adamson, ORNL
		• Department of Defense High Performance Computing Modernization Program (DoD HPCMP) Overview
		  o Stephen Bowman, Rickey Gregg, and Gary Key, HPCMP
		• Keeping It All Safe: LLNL HPC Security Architecture
		  o Ian Lee, LLNL

11:45 a.m.	1:00 p.m.	Lunch
1:00 p.m.	2:15 p.m.	HPC Operator Security Experience • Jared Baker, NCAR • Nathaniel Mendoza, TACC • John Walker, NCSA/UIUC • Scott Sakai, SDSC/UCSD • Andrew Adams, PSC
2:15 p.m.	2:30 p.m.	Break

2:30 p.m.	3:45 p.m.	RMF Development, Implementation, and Assessment
		• Chair: Vicky Pillitteri, NIST
		• The Trusted CI Framework
		  o Jim Basney, UIUC
		• Development of the TOSS 4 STIG
		  o Ian Lee, LLNL
		• Panelists:
		  o Jennifer Allphin, Jim Waterman, Steve Boettcher, and Gary Key (HPCMP)
		  o Aron Warren, SNL
		  o Catherine Hinton, LANL

3:45 p.m.	5:00 p.m.	Breakout Sessions
		• HPC System Vulnerabilities and Threats
		  o Ryan Adamson (Lead), Csilla Farkas (Scribe)
		• HPC RMF (Risk Management Framework): Challenges and Opportunities
		  o Erik Deumens (Lead), Aron Warren (Scribe)
		• HPC Security Implementations, Best Practices, and Challenges
		  o Catherine Hinton (Lead), Gary Key (Scribe)
		• Future HPC System and Its Implications to Security
		  o Albert Reuther (Lead), Puri Bangalore (Scribe)

Table 2. 3rd High-Performance Computing Security Workshop Agenda — Day 2

Start Time (ET)	End (ET)	Session/Talk
8:45 a.m.	9:30 a.m.	Keynote 2: Office of Science HPC Cybersecurity • Dr. Robinson Pino, DoE

9:30 a.m.	10:20 a.m.	HPC Security Research
		• Chair: Dr. Xin Yuan
		• CryptMPI: A Fast Encrypted MPI Library
		  o Dr. Xin Yuan, FSU
		• Automated Detection of Configuration-Based Vulnerabilities in HPC Workload Managers
		  o Dr. Mu Zhang, University of Utah

Start Time (ET)	End (ET)	Session/Talk
		• Searchable Encryption on Scientific Data o Dr. Hoda Maleki, Augusta University
10:20 a.m.	10:40 a.m.	Break

10:40 a.m.	11:30 a.m.	• Regulated Research Community of Practice
		  o Erik Deumens, UFL
		• Security and Accelerators: A Research Perspective
		  o Joseph Manzano and Nick Multari, PNNL
		• HPC Provenance and Security
		  o Yong Chen, TTU

11:30 a.m.	12:15 p.m.	Keynote 3: Usable Computer Security and Privacy to Enable Data Sharing in High-Performance Computing Environments • Dr. Sean Peisert LBNL
12:15 p.m.	1:15 p.m.	Lunch

1:15 p.m.	2:45 p.m.	HPC Vendors
		• Chair: Sanjay Rekhi, NIST
		• Data center-wide security by default
		  o David Reber, Nvidia
		• A Brief Introduction to Security Primitives in the Kubernetes Ecosystem
		  o Jeremy Duckworth, HPE
		• Beyond the Walled Garden: Cloud Security Patterns for HPC
		  o Lowell Wofford, AWS

2:45 p.m.	2:55 p.m.	Break
2:55 p.m.	3:35 p.m.	Breakout Readout
3:35 p.m.	3:50 p.m.	HPC Security Working Group Readout • Yang Guo, NIST
3:50 p.m.	4:00 p.m.	Wrap Up

2. Workshop Session Highlights and Summaries

2.1. HPC Architecture and Security Posture

High-performance computing presents distinct security challenges, as it requires striking a balance between achieving enhanced performance and maintaining system security. This session primarily focused on the interplay between performance and security. Presenters from government agencies and national laboratories shared their work on securing HPC systems and offered valuable insights into the ongoing challenges faced by their respective organizations. It is noteworthy that these challenges encompass numerous environments and security requirements.

Andrew Prout from Massachusetts Institute of Technology (MIT) Lincoln Lab discussed their use of virtual private HPC as a software development sandbox. HPC systems are utilized for a wide range of self-developed software that serves specific research purposes. Unfortunately, security considerations are often overlooked in the development of such software. Andrew discussed potential techniques that can effectively manage and address the security implications associated with these workflows. One of the techniques highlighted was the implementation of virtual private HPC, which creates a clear separation among users. This separation includes limiting users' visibility into the activities of others, such as processes, filesystem access, network traffic, and accelerator usage. With this setup, individual users experience a sense of running on their own dedicated HPC system, thus enhancing both security and privacy.

In the second presentation, Aron Warren from Sandia National Laboratories introduced the concept of a diskless HPC architecture. This innovative approach aims to minimize operational overhead while offering multiple advantages for enhancing security. By eliminating physical disks at most systems, it reduces the risk of data breaches or unauthorized access to sensitive information. Additionally, the diskless setup streamlines maintenance processes, reduces the attack surface, and enhances overall system resilience.

According to Ryan Adamson from the Oak Ridge Leadership Computing Facility, the design and architecture of HPC systems severely lack security considerations. These systems prioritize speed and performance, often relegating security to an afterthought. In his presentation, Ryan examined the security-related disparities between industry-standard commodity systems and the common characteristics found in HPC environments. He delved into the interfaces that connect HPC components — particularly those utilized for HPC workflows — to evaluate the strengths and weaknesses of the existing HPC security model. Ryan also provided an example of a zero-day vulnerability that was discovered in a system formerly ranked in the TOP500 list. The demonstration emphasized the need for broader action in implementing robust security measures within the HPC field.

The Department of Defense High Performance Computing Modernization Program (DoD HPCMP) is entrusted with the mission of accelerating technology development and the transition toward superior defense capabilities through the strategic utilization of high-performance computing, networking, and computational expertise. Stephen Bowman, Rickey Gregg, and Gary Key described the mission of the DoD HPCMP and outlined their security strategies for maintaining multi-tiered cybersecurity measures and risk management processes to minimize risk and enhance the resiliency of HPCMP networks and systems.

During the final presentation, Ian Lee from Lawrence Livermore National Laboratory (LLNL) provided an overview of the security operations conducted at Livermore Computing. As the complexity of HPC systems continues to grow to meet user needs, the challenges associated with securing these systems have also escalated. While security compliance frameworks offer some assistance, they are not the sole solution to ensuring comprehensive security. Ian emphasized the critical importance of system monitoring, gaining insights into ongoing activities, and illuminating the computing environment. He also described the implementation of a continuous monitoring system at LLNL, which effectively addresses these requirements. This system enables proactive monitoring and offers valuable visibility into the computing environment to enhance security measures.

2.2. HPC Operator Security Experience

HPC system administrators play a critical yet often under-appreciated role in maintaining the availability, performance, and security of HPC systems and ensuring that they serve their mission as a powerful computational instrument that enables scientists and science. These HPC operators often work behind the scenes to meet the daily operational needs and challenges of running the infrastructure, supporting users and software applications, and responding to security events.

As system administrators bridge the gap between the theory and practice of securing HPC systems, the workshop organizers were delighted to convene a panel of five operators of NSF-supported HPC with representation from National Center for Atmospheric Research (NCAR), National Center for Supercomputing Applications (NCSA), Pittsburgh Supercomputing Center (PSC), San Diego Supercomputer Center (SDSC), and Texas Advanced Computing Center (TACC). The goal of the panel was to facilitate an open exchange between practitioners, security researchers, and policy and standards makers in order to identify best practices, common challenges, and anticipated future threats.

These HPC systems serve a primarily academic user base who engage in open science. NSF does not mandate a specific cybersecurity compliance regime for the HPC systems that it helps to support but recognizes that each facility may be a unique environment with unique requirements that demand an approach tailored to its unique mission. The science and workloads of each HPC system can vary greatly. For example, NCAR focuses on climatology models and atmospheric research, so the panel's responses reflected the diversity of their HPC missions, users, and approaches to cybersecurity.

The panelists articulated some of the unique aspects of their HPC systems that inform their approach to security. Some of these differences included the use of dedicated login (bastion) nodes, the fact that users often do not belong to the same organization that owns or runs the HPC, the highly dynamic nature of the workloads, and the fact that users can bring and compile their own code. One panelist described how even the well-accepted practice of using multi-factor authentication is challenging to accomplish with off-premises users who do not belong to the organization.

Another panelist described how the traditional use of login or stepping-stone nodes has evolved, with users now demanding new tools and interfaces instead of the traditional Secure Shell Protocol (SSH). This trend toward new and more featureful access methodologies with more

software dependencies and a larger attack surface has presented the operators with new challenges in ensuring that users can operate securely.

Indeed, users themselves can present threats to HPC systems. In an HPC environment, it is not unusual for users to bring their own code and compile their own binaries. The panel agreed that this open software policy is necessary to support the unique needs of their diverse population of science users but also agreed that user-installed software presents challenges to understanding the threat surface and how to identify and patch vulnerable code. The panel unanimously agreed that they strove to support users and their unique application needs but stressed that they have to be aware of these needs and that users should engage earlier and more deeply with the operators. The panel also discussed how the inherently multi-node, distributed nature of typical HPC systems presents an advantage in performing rolling reboots, wherein machines are patched when their jobs are finished. As the job scheduler will simply choose an available node, the patch and reboot are transparent to the end users.

The panel identified software supply chain security as both a current and future challenge, especially in HPC, where code provenance is largely unknown and users compile their own code. Asked how they might redesign HPC systems to be more secure, one panelist emphasized the power of virtualization, containers, and the ability to limit the attack surface. The panel discussed the relative merits, performance, and obstacles to such a pure container-based approach.

In contrast, one panelist discussed how they have embraced end users as a second set of eyes or a "human intrusion detection system." The panelist noted that since the users are most intimately familiar with their workload and how it performs on the HPC system, they are often able to identify problems that could be indicative of a system failure or a larger security event that warrants investigation. This was another example of how deeply engaging with HPC users yielded benefits for this particular operator, wherein their users alerted them to potential anomalies before the operators themselves were aware of the problem.

In the end, the panel enthusiastically identified with the HPC mission and emphasized that the "supercomputer is a research instrument." They further highlighted the market-like forces that incentivize them to provide particular capabilities, both in terms of functionality and security. If users are unable to effectively use one HPC system, they may choose to use a different one. Thus, the operators concluded that making HPC performant, featureful, and secure ultimately supports users, impacts science and discovery, and ensures a continued stream of funding.

Finally, when queried about their hypothetical primary concern in the context of a cybersecurity incident ("what keeps you up at night"), the panel agreed that the most worrisome potential consequence is that the science itself is somehow called into question. The panelists emphasized how this fear has driven their continued work to secure their respective HPC and to continue to balance the security and usability of the system amid a continually evolving threat and policy landscape.

2.3. Risk Management Framework Development, Implementation, and Assessment

The Risk Management Framework (RMF) Development, Implementation, and Assessment Panel featured panelists from federal agencies, national laboratories, and academia, all of whom extensively utilize HPC to support operational research missions. The goal of the panel was to

discuss how the NIST RMF supports their HPC cybersecurity programs and share valuable insights, best practices, future challenges and opportunities, and lessons learned. Two panelists presented their work on developing a security framework for scientific HPC facilities and supporting the implementation of RMF, and a moderated discussion to exchange ideas followed.

2.3.1. Presentation on the Trusted CI Framework

Dr. Jim Basney from the National Center for Supercomputing Applications (NCSA) at University of Illinois Urbana-Champaign (UIUC) presented an overview of the Trusted Cyberinfrastructure (CI) Program. The program aims to establish a comprehensive framework to address cybersecurity and privacy needs within the scientific community. HPC facilities supported by the NSF are crucial for NSF research projects. NSF HPC facilities are largely responsible for their own cybersecurity and are not required to follow existing federally legislated security and privacy requirements.

The NSF's security approach avoids prescribing specific security measures, enabling facilities to strike a balance between cybersecurity requirements, unique research objectives, and the constraints of the science mission. Trusted CI has developed the "Trusted CI Framework" as a set of guiding security policy pillars for NSF projects and facilities. The Trusted CI Framework pillars serve as the foundation for customized security and privacy requirements tailored to the unique needs of research facilities and their scientific missions.

Furthermore, the structured Trusted CI Framework can be seamlessly integrated with existing federal security and privacy guidance. Currently, a crosswalk is being developed to showcase the alignment between the Trusted CI Framework and the NIST RMF. Additional information can be found at the Trusted CI Framework website.

2.3.2. Presentation on the Development of TOSS 4 STIG

Ian Lee from Lawrence Livermore National Laboratory (LLNL) introduced the development process of a publicly available Security Technical Implementation Guide (STIG) for the Tri-Lab Operating System Stack (TOSS) HPC environments. Starting with the identification of a common operating system for HPC environments, the team focused on building a methodology that addressed the custom requirements needed for hardware architectures, high-speed interconnections, and very large filesystems.

Recognizing the challenge of tailoring STIGs for each implementation, the project sought to establish a standard that could be leveraged for other HPC environments. Starting with the Red Hat Enterprise Linux 8 STIG, the team adjusted STIG-specific language to enhance its relevance to other government agencies. Close collaboration with the Defense Information Systems Agency (DISA) ensured that requirements are included to meet HPC-specific operational requirements. The rules in the TOSS4 STIG include additional information about the HPC implementation as well as check and fix elements that enable users to implement their own validation methods. The STIG is available on the Cyber Exchange public site. Additional information can be found on the project's GitHub page.

2.3.3. Panel Discussion

The session moderator opened the panel discussion with a question regarding the application of the NIST RMF in the panelist organizations' HPC environments. The responses highlighted two distinct strategies:

- **Organization-directed strategy:** In this approach, the organizational security program determines the requirements applied for all information systems within the organization. Panelists discussed the challenges of aligning the organizational security program with the unique technical characteristics of HPC. The ongoing challenge has been to find a viable solution to address all of the controls mandated by the organization and manage organization-specific risks. This frequently involves the identification of compensating controls to tackle implementations that may be deemed "non-compliant" relative to other enterprise systems and components.
- **Project-directed strategy:** Conversely, other organizations recognize the unique security requirements of HPC and reap the benefits of having security management within the HPC project team. This approach is more flexible and caters to the unique requirements of HPC missions and users.

The methods used to determine the authorization boundary for an HPC environment vary among the panelists' home organizations. In certain organizations, the authorization boundary is based on the location or site, resulting in the HPC system being placed in the same authorization boundary as other types of systems. In contrast, other organizations isolate HPC systems within a distinct authorization boundary.

Tailoring RMF implementation for HPC has required a certain level of "creativity" in the selection of controls and the documentation of control implementation details. In organization-directed implementations, the controls are selected at the organization level using the NIST Special Publication (SP) 800-53B control baselines. Further tailoring is often required to address unique HPC operational requirements. Environments that are more project- or mission-oriented have more flexibility with the application of the baselines and the selection and tailoring of SP 800-53 controls. As expressed in the presentation about the Trusted CI Framework Project, the science community has avoided a prescriptive approach to security.

The discussion delved into the integration of HPC with organization and enterprise risk management policies, as well as the importance of communication between front-line operations and organization and enterprise management. While HPC can be regarded as a distinct enclave within the organization, any integration with broad risk management policies should recognize the difference and uniqueness of HPC in comparison with other user operations.

The panelists agreed that risk assessments for HPC environments require both continuous evaluation and constant communication between system stakeholders and organization management. The collection and evaluation of compliance metrics and performance metrics provide ongoing insights into whether standards are being met and whether improvements are necessary. This is particularly significant for maintaining the balance between security configurations and operational performance. When HPC systems can be distinguished from the rest of the enterprise, an HPC-specific control overlay may be more meaningful in determining distinct HPC security requirements.

Specific security requirements, such as auditing services, has the potential to impact the availability of computing resources. This issue has been identified as a common challenge across HPC environments. Another key challenge and opportunity is the need for effective communication and information sharing (e.g., sharing lessons learned). In many cases, organizations go through their own experimentation process on a case-by-case basis to derive local solutions that address requirements or system issues. For instance, one organization has developed a solution for limiting the performance impact of the auditing service. Sharing that information through a common knowledge repository would be beneficial across the HPC community.

The panelists agreed that communicating HPC challenges and needs to organizational management requires the ability to express the risks without relying on cybersecurity jargon that may not be universally understood.

There is also a growing concern that the security talent pool is shrinking, and outreach efforts are needed to bring in the next generation of professionals. The existing community is already facing limitations and experiencing attrition like other sectors. Creating more recruiting opportunities and building on the existing knowledge base will be vital to the ongoing success of HPC security in the future.

In summary, the panelists' recommendations emphasized the importance of comprehending the HPC operational environment. This includes documenting security and privacy requirements and keeping track of security and operational performance metrics, which will allow management to gain valuable insights into how well the HPC is performing in terms of both security and mission objectives. Furthermore, automation will help achieve consistent implementations for standardized and objective requirements. Tools like Gitlab and Ansible play a significant role in facilitating the maintenance and sharing of scripts that support the consistent implementation of security requirements unique to HPC environments.

2.4. HPC Security Research

In the HPC security research session, six projects in different areas and varying maturity stages were presented:

1. Dr. Xin Yuan from Florida State University presented the research and development of an encrypted Message Passing Interface (MPI) library called CryptMPI. Encryption and decryption operations are computationally intensive. To achieve high encrypted communication performance, CryptMPI has incorporated novel protocols and algorithms. It has also been evaluated against other software encryption solutions in container clusters, such as Docker Swarm. The results show significant performance advantages to encrypting messages at the library level over solutions that encrypt messages at the system level.
2. Dr. Mu Zhang from the University of Utah presented research on automated detection of configuration-based vulnerabilities in HPC workload managers. Configuration-based vulnerabilities in other contexts, such as the mobile environment, have been extensively investigated. In the HPC domain, however, there has been little prior research despite configuration-based vulnerabilities becoming a serious problem. This nascent project started from the original idea of applying techniques developed in other contexts to the

HPC domain but has grown into a larger project of building a general infrastructure that supports automated detection of vulnerabilities.

3. Dr. Hoda Maleki from Augusta University presented research on searchable encryption of scientific data. Searchable encryption is a popular field, and many techniques have been developed, largely for cloud computing environments. This research focuses on developing searchable encryption specifically for scientific data, which have distinct characteristics from other types of data. The project has just started, and Hoda mostly discussed what is to be expected in this project.
4. Dr. Erik Deumens, Director of University of Florida Information Technology Research Computing, gave an informative talk on their efforts to create a community of practice for regulated research. He discussed the need for standardization in this domain and the SP 800-223 HPC security effort. He also described activities to build the community, including monthly webinars, sharing resources, and organizing workshops. Additional information can be found at the Regulated Research Community of Practice (RRCoP) project website.
5. Dr. Joseph Manzano from Pacific Northwest National Laboratory (PNNL) gave a talk on security issues for systems with accelerators. He described a scenario in which the accelerator hardware can be exploited to infer information.
6. Dr. Yong Chen from Texas Tech University presented the research and development of a software infrastructure that provides lightweight provenance and a service for the collection, management, and analysis of provenance data (i.e., the metadata that describes the history of the data). He also discussed how such a system can be used for security purposes. The system has been deployed at the computing center at Texas Tech, and Dr. Chen gave a demonstration of the system in the presentation. The capability of the infrastructure can improve the productivity of science in complex HPC simulation and analysis.

2.5. HPC Vendor Viewpoints

Three speakers from different vendors presented their views:

- Jeremy Duckworth from Hewlett Packard Enterprise (HPE) spoke about securing the Kubernetes ecosystem, which has been adopted as their HPC management system. He began by highlighting three aspects of Kubernetes security: runtime security, name-space-level network security, and service mesh security. He further explored their application to the Kubernetes-based Cray HPC management system and explained the Kubernetes pod execution model and Kubernetes runtime policy engine, which supports dynamic admission control for application programming interface (API) requests. As an example, a Kubernetes manifest is verified or forced to make changes to adhere to security requirements, and container images are verified before execution. Jeremy proceeded to discuss Cilium-based Kubernetes network policy enforcement for network communications between name spaces and Istio-based service mesh security, which automates various application network tasks in a transparent and language-independent manner. Finally, Jeremy presented the Cray system management architecture that

manages the entire HPC system via multiple service meshes, networks, and security policy enforcement points.

- David Reber from Nvidia delivered a presentation titled "Data Center-Wide Security by Default," which tackled the challenges that stem from the disaggregated corporate perimeter and distributed computing, including the user-edge-central data-center computing paradigm and federated model training. Nvidia utilizes confidential computing (CC), Morpheus, and data processing units (DPUs) to enhance security. CC establishes a trusted computing environment that facilitates secure deployments of proprietary artificial intelligence (AI) models and federated learning. NVIDIA Morpheus — a software development kit (SDK) accelerated by graphics processing units (GPUs) — empowers cybersecurity developers to construct optimized AI pipelines for real-time data filtering, processing, and classification. Additionally, DPUs represent a novel class of programmable processors that offer services such as compute attestation, storage file management, network pluggable alternatives, and telemetry. Toward the conclusion of the talk, David emphasized the importance of adopting innovative technologies like DPUs, CC, and Morpheus to ensure secure HPC in the future.
- Lowell Wofford from Amazon Web Services (AWS) presented his perspective on HPC security in a talk titled "Beyond the Walled Garden: Cloud Security Patterns for HPC." As HPC computing expands beyond a single trusted computing zone, securely orchestrating workflows becomes a significant challenge due to numerous interactions among different components. Creating custom security controls for each interaction is neither scalable nor reliable. To address the issue, one potential solution is to adopt an API-driven design with unified APIs for the entire computing infrastructure. By enforcing API access policies at the unified API layer, access can be unified and streamlined. The second challenge explored in this presentation was the dynamic creation of trusted computing zones for individual tenants or applications in a multi-tenant, distributed computing infrastructure. Lowell suggested that such requirements can be facilitated by making trust programmatic and declarative so that individual walled gardens of trust zones can be created on demand.

3. Breakout Session Key Takeaways

3.1. HPC System Vulnerabilities and Threats

The attendees of this breakout session engaged in discussions aimed at identifying vulnerabilities, threats, and threat actors that are significantly distinct from an enterprise information technology (IT) environment.

The first question focused on identifying the threat actors who target HPC systems and understanding how they differ from those who target typical personal, enterprise, or government IT systems. The participants of the breakout session identified four categories of threat actors, ranging from adventure-seeking students to sophisticated nation-state actors:

1. Authorized users who make unintended errors versus attackers who intentionally carry out an attack
2. Attackers with specific goals versus attackers who simply want to create havoc
3. Novice attackers versus attackers with knowledge of HPC
4. External attackers versus authorized users who push the boundaries of the system

Considering the attackers' motives is crucial to assessing the potential damage that they can inflict on an HPC system. A novice attacker with limited skills may aim to gain unauthorized access to the HPC system but pose minimal risk of significant or sustained harm. Conversely, a disgruntled insider with access to all facets of the HPC system could potentially cause extensive and enduring corruption across multiple projects. The participants of the breakout session discussed the following potential motives for intentional attacks:

- Attacks against availability by exploiting vulnerabilities in HPC components (e.g., software, firmware, hardware, and network protocols) with the aim of rendering the system unavailable
- Attacks against confidentiality through unauthorized access to proprietary data and metadata
- Attacks against integrity (data-driven) by manipulating data and/or metadata to cause the HPC system to produce inaccurate results or to achieve specific outcomes intended by the attacker
- Attacks against authenticity by exploiting weak security tools, practices, and human errors to gain unauthorized access to HPC resources

The second working group discussed the unique threats and vulnerabilities associated with HPC systems. The participants of the breakout session identified the following distinct HPC vulnerabilities:

- Most vulnerability researchers and offensive security experts are not familiar with HPC systems.
- Users bring their own software that may be unpatched and may pose security risks.
- HPC systems may have to interface with older or unpatched software, increasing the potential for vulnerabilities.

- System management services may expose specific vulnerabilities that could be exploited by attackers.
- HPC systems support multi-user simultaneous use of the system without complete segmentation.
- Securing high-speed interconnects in HPC systems requires specific considerations that differ from other distributed systems.

The third working group discussion point focused on evaluating the various impacts of potential cyber attacks from the perspectives of various HPC stakeholders (e.g., system owners, vendors, end users, and the scientific community), especially impacts to scientific discovery, business decision making, and reputation. The group agreed measuring such impacts may encompass multiple factors, such as financial loss, reputation loss, and environmental impacts.

The fourth working group discussed the security measures utilized by enterprise information systems and their effectiveness when applied to HPC systems. The attendees specifically explored the measures that are most and least effective when implemented in HPC systems. The consensus was that the following security best practices are often straightforward to implement and highly effective in an HPC environment:

- Multi-factor authentication
- Network security monitoring
- Intrusion detection
- Vulnerability scanning
- Log aggregation
- Antivirus scanning
- Endpoint detection
- Hardware-based encryption

The least effective security controls, which also significantly impact system performance, were identified as:

- Software-based encryption
- Patching and software maintenance
- Endpoint protection
- Anti-virus software
- Signature-based detection of malicious software

3.2. HPC RMF: Challenges and Opportunities

This session focused on discussing the challenges posed by the SP 800-53 security controls that are the most difficult to implement on HPC systems:

- Antivirus controls (SI-3): The requirement to scan Petabyte-scale file systems takes an extensive amount of time and leads to significant performance issues for storage operations. Additionally, some antivirus tools do not offer the option to exclude filesystems or coordinate multiple scanners, resulting in multiple simultaneous scans of the same filesystem. Targeting scans on the edge and gateway nodes, where files enter the HPC system, as well as the file systems that hold the system and configuration data should address the highest risk.
- File integrity monitoring (SI-7): Configuration management and validation can be implemented using modern tools instead, such as Puppet Check.
- Logging of all file accesses (AU-2, AU-6): In particular, all read accesses cause performance degradation on large parallel file systems with high I/O (input/output) activity. One solution is to provide compensation information to identify the files that could be accessed without checking each file (e.g., monitor at the directory level).
- Intrusion detection checking causes performance problems (SI-4): Using out-of-band intrusion detection provides the needed visibility into the data flow without impacting performance. Walking the filesystem can also determine whether files have changed.
- Scanning all open ports, protocols, and services (CM-7(1)): It is challenging to scan all open ports, protocols, and services in HPC development environments due to the large amount of application development activities. Instead, the daemons, services, and ports that support the cluster should be listed as they are known services. Any services or daemons created by users to be "published" also need to be reviewed. In addition, scanning edge nodes provides needed visibility of the greatest risks.
- Application allowlist/denylist (CM-7(5)): This control is infeasible in a development environment. Most applications have a simple data flow for input and output, are heavily compute bound, and are executed in non-privileged mode. Allowlisting or denylisting does not change the risk profile significantly. The applications that involve ports to the outside network need closer scrutiny before deployment.
- Static code analysis and rigorous code review required for researching software development causes unacceptable delays in the software development process (RA-5): A software process where code goes through stages of development, testing, and production allows for effective risk management without delaying the development cycle.
- Unsupported hardware and software are common on research HPC systems (SA-22): There is a time-delay gap between when vendors provide novel HPC systems and system components and when authorization officers are comfortable approving them.

The session also discussed other security control issues and the best practices that have been implemented in practice:

- STIG requirements can pose implementation problems. The STIG may call out specific vendor products instead of focusing on desired functionality. Ideally, the STIG should be designed similarly to the RMF and allow sites to partially implement it, just as SP 800-53 allows partial implementation of a control with proper justification and risk analysis. The STIG should not be an all-or-nothing approach. Additionally, the STIG often contains

specific values, such as 10 concurrent sessions, and should instead allow sites and organizations to choose the values.

- Promote the definition of HPC as a single, monolithic system. It should not be considered a collection of a thousand different systems (e.g., comprising compute nodes, admin nodes).
- Vulnerability mitigation (SI-2) timeframes are difficult to follow. Jobs may have been running or waiting in the queue for a long period of time, and patching will terminate those jobs. Rolling patches should be encouraged whenever possible.
- There are no good tools for identifying where software vulnerabilities reside.
- The SP 800-53 overlay should have more "organizationally defined value" instead of specific values, which will allow for a tailored value based upon risk, type of vulnerability, and other factors. For example, a High-Moderate-Low should have a 60-day mitigation period for a critical vulnerability, but a Moderate-Low-Low could have a 90-day mitigation period for critical vulnerabilities. The Committee on National Security Systems Instruction (CNSSI) gives some guidance on scoping organization-based parameters. In the HPC security overlay, the overlay should give guidance on what the values should be and explanations for why certain values are possible or needed.
- There are broad concerns about disk encryption and encryption at rest. Physical security and the striping of data across many drives mitigate the risk of data loss by limiting physical access to disk drives.
- There were general questions about the Federal Information Processing Standards (FIPS) 140-2 and the effects of post-quantum-resistant algorithms on systems.
- Authorization boundaries should be investigated more closely in system security plans (SSPs).
- Maintaining documentation in the verbiage and terminology of RMF regarding where one control may be tailored differently from another and/or contradict verbiage in other security controls would benefit the SSP.

3.3. HPC Security Implementations, Best Practices, and Challenges

This working group session discussed security standards and best practices that HPC security staff are either currently implementing or would like guidance on how to implement. There were approximately 18 attendees (system administrators, policy makers, and program integrators) who represented various industries, including academia, healthcare, and government. The conversations had a similar theme: applying security standards designed for single server instances across multi-node HPC or cluster supercomputers is difficult.

The group discussed a few areas of concern. One area was architecture understanding — where certain controls and configurations need to be applied (i.e., which node type) and the specific configurations that can be used to meet requirements. Another area was organization adoption or user acceptance. HPC is commonly used for scientific research and development, which require speed, agility, and flexibility in cross-node communication, storage transfers, and parallel compute capabilities, all of which could be hindered by overly restrictive security controls.

There were several comments about security implementation challenges within the HPC environment. Some of the concerns centered on the education of IT personnel, who may lack training on the unique nature of HPC systems and their use. Another area of concern involved trying to apply the same rules to HPC assets as enterprise-type systems, most of which currently follow non-HPC guidance (i.e., SP 800-171 or SP 800-53) in lieu of HPC-specific standards and best practices.

There were comments about the conflict between system administrators and security and the difficulty of finding comment ground between the two groups regarding HPC. One suggestion was to include input from both system administrators and security groups in acceptance standard operating procedures (SOPs), policies, best practices, and even common dashboards to help bridge the gap.

The group also discussed the common challenge of testing security implementations. All agreed on the need for viable test environments for function and performance benchmarking, which are often cost prohibitive. As a result, many have utilized smaller scale clusters, virtual alternatives, and zoned system implementations within an existing cluster. It was noted that at least some testing could be performed using these methods.

3.4. Future HPC System and Its Implications for Security

The discussion on Future HPC Systems was structured into three key areas: hardware, software, and policy.

3.4.1. Hardware

- Composable/disaggregated systems (CXL) and their implications on architectures
- Multi-tenancy: Exploring how to efficiently share resources among multiple users or applications
- Supply chain security: Ensuring the integrity and security of hardware components throughout the supply chain
- Real-time authorization of components: Enabling the verification and authorization of components in real time
- Transition from Trusted Platform Module 2.0 (TPM2) to Virtual TPM (VTPM): Evaluating the benefits and challenges of transitioning from TPM2 to Virtual TPM
- SmartNICs and the software executed on them: Understanding the role of SmartNICs and the software running on them for enhancing system performance and security
- Software-defined networks (SDNs): Leveraging SDNs to enable more flexible and efficient network management

3.4.2. Software

- Crypto agility: Developing software that can adapt to different cryptographic algorithms and protocols to ensure long-term security

- Virtualization and containerization: Exploring the use of virtualization and containerization technologies to improve system scalability, isolation, and manageability
- Designing secure software and vetting processes: Implementing robust software development practices and thorough vetting processes to minimize security vulnerabilities
- Signed software libraries and components: Ensuring the authenticity and integrity of software libraries and components through digital signatures
- Support of legacy code with minimal attention to security: Addressing the security concerns associated with legacy code to protect against vulnerabilities

3.4.3. Policy

- Transition to RMF and the subsequent move to the Cybersecurity Maturity Model Certification (CMMC) in the coming decade
- Zero trust approach and its relevance to HPC systems: Implementing a zero trust architecture that assumes no inherent trust in the system components and networks
- Supply chain management: Developing strategies to mitigate the risks associated with the supply chain, including hardware and software components

These discussions highlighted the importance of considering hardware, software, and policy aspects when designing and operating future HPC systems. By addressing these areas, stakeholders can work toward building more secure, efficient, and resilient HPC environments.

4. Keynote Summary

4.1. Keynote 1 — The NSF HPC Security Landscape: Research Challenges to Production Capabilities

The NSF has long played an important role in supporting HPC and the ecosystem of surrounding data, networking, software, and advanced computation technologies. Dr. Robert Beverly from the NSF's Office of Advanced Cyberinfrastructure (OAC) outlined OAC's current landscape of HPC cyberinfrastructure and outstanding challenges related to HPC security. He then provided insights into NSF's strategic direction in HPC cybersecurity and a forward-looking vision for the future of HPC security.

OAC's mission is to support and coordinate the development, acquisition, and provisioning of state-of-the-art cyberinfrastructure resources, tools, and services that are essential to the advancement and transformation of science and engineering. OAC cyberinfrastructure is designed to enable research and researchers across the range of core scientific disciplines that the NSF supports. Dr. Beverly emphasized the holistic view that OAC takes with cyberinfrastructure to facilitate scientific discovery and innovation, including advanced computing, data, software, networking, and people and workforce development.

OAC's current portfolio of cyberinfrastructure includes leadership-class HPC, as well as experimental computing architectures, cloud resources, campus interconnection, testbeds, and coordination services. OAC is a firm proponent of open science and recognizes the unique environments, instruments, users, and experiments within the NSF community. In particular, the NSF does not seek to mandate a particular security compliance regime or to perform security audits. In essence, NSF does not want cybersecurity to impede science or scientists but, rather, to be an enabler of science.

In a world of open science, Dr. Beverly explained that cybersecurity can still play an important role. For instance, he articulated a future in which:

- Data have strong integrity protection to protect against accidental or malicious modification.
- Research artifacts contain provenance metadata.
- Collaboration between scientists and infrastructure is seamless and natural.
- Computation and the sharing of sensitive data (e.g., personally identifiable information (PII)) are possible without compromising privacy.
- Infrastructure has high availability and is not vulnerable to misuse.
- Third parties can replicate and reproduce research findings.
- The public trusts science.

To this end, OAC's vision involves tightly integrating cybersecurity into computing, network, and data to provide an agile, robust, trustworthy, and sustainable cyberinfrastructure ecosystem. In this fashion, the infrastructure should be collaborative, accessible, and reproducible to enable science and discovery.

Dr. Beverly described some of the ways in which HPC security is unique and presents distinct challenges:

- May involve novel architectures, such as neuromorphic and quantum
- Typically involves specialized software, workloads, and data
- Has a unique population of users
- Focuses on performance
- Has a distinct science mission
- Has different adversaries than, for example, enterprise computing

OAC and the NSF have taken a holistic approach to securing science cyberinfrastructure through multiple programs that target different capabilities and challenges. For instance, the Secure and Trustworthy Computing (SaTC) program engages in fundamental cybersecurity research, while the Cybersecurity Innovation for Cyberinfrastructure (CICI) program targets applied and translational cybersecurity work for science cyberinfrastructure. CICI has three program areas: usable and collaborative security for science, reference scientific security datasets, and a transition to cyberinfrastructure resilience track. Across these areas, Dr. Beverly highlighted some particularly exciting projects, including work on improving the performance and security of data transport, analyzing and automatically patching scientific binary software, finding and fixing configuration vulnerabilities in HPC, performing encrypted search and computing on private data, and distributed credentialed access.

Dr. Beverly concluded by speaking about OAC's initiatives for supporting the human side of cyberinfrastructure through learning and workforce development, cyberinfrastructure professionals, and the minority-serving cyberinfrastructure consortium.

4.2. Keynote 2 — DoE's Office of Science HPC Cybersecurity

Dr. Robinson Pino, a program manager at the Department of Energy's (DoE) Office of Science, delivered a keynote address on the scientific programs and associated activities that center on HPC security at the DoE Office of Science. As the primary agency responsible for facilitating scientific breakthroughs and developing major scientific tools, the Office of Science plays a pivotal role in transforming our understanding of nature and advancing the energy, economic, and national security of the United States. It is the largest supporter of physical sciences in the U.S. and provides funding to over 300 research institutions, including 10 national laboratories.

The HPC facilities at the Office of Science, such as the top-ranked Frontier system, are open to all interested users and offer free access if the users publish their research results in the open literature. The Office of Science manages a research portfolio with a key focus on advanced scientific computing. To fulfill its mission, the DoE must ensure the integrity and availability of scientific facilities, software, and data. The Office of Science operates in open environments, and its HPC and large-scale science workflows differ from general-purpose computing. For instance, DoE machines may run one program for weeks on tens of thousands of processors. The DoE has actively supported various HPC cybersecurity research, including trustworthy supercomputing, high-end networking, data centers, and research to detect attack patterns by correlating

heterogeneous sources of information, such as the network, computing nodes, operating system, runtime, and applications.

Dr. Pino also discussed two new scientific initiatives that are closely related to HPC. The first initiative, titled "5G for Science," explores the potential of emerging 5G wireless technologies. The new technologies offer opportunities and capabilities to enable the digital continuum that links the wireless edge to advanced scientific user facilities, such as HPC. The second initiative focuses on microelectronics research and development (R&D), which holds the promise of revolutionizing memory and data storage to redefine the future computing.

4.3. Keynote 3 — Usable Computer Security and Privacy to Enable Data Sharing in High-Performance Computing Environments

Dr. Sean Peisert from the Lawrence Berkeley National Laboratory (LBNL) delivered an informative and engaging keynote speech on enabling technologies for securing data sharing in HPC systems. Data play a critical role in scientific computing, particularly with the increasing importance of AI and Machine Learning (ML) model training. However, sharing data while preserving confidentiality, integrity, and availability presents significant challenges. Regulatory requirements, the proprietary nature of data (such as trade secrets), and the inclusion of private and sensitive information in data are major hurdles to data sharing for scientific advancement. Dr. Peisert began by reviewing current data-sharing practices and delved into two novel technologies that protect the data from untrusted end-users and computing platforms: differential privacy (DP) and trusted execution environments (TEEs).

Physical isolation is a common technique for protecting data from external threats, but it is ineffective against internal threats posed by users and HPC providers. Another approach is data anonymization, which removes personally identifiable information from data sets to protect privacy. However, as the volume and diversity of available data increase, anonymization can be defeated. Differential privacy offers a solution by maximizing analysis accuracy of sensitive data while minimizing the risk of individual re-identification. This is accomplished by adding Laplace or Gaussian noise to statistical database query responses within a predetermined range. DP has already been adopted by Apple, Google, and the U.S. Census Bureau to protect data privacy and is ready to be employed in HPC security settings. Dr. Peisert's work on using DP to anonymize grid data for cyber attack detection and enable privacy in mobility data, including aggregated mobility networks and individual vehicle trajectories, further demonstrates the approach's effectiveness.

TEEs allow data owners to share data without placing trust in the compute facility. TEEs implement hardware-based process isolation and encryption of memory and computation, which effectively prevent data exposure to other users and even system administrators. Additionally, they can provide cryptographic attestation to verify execution. To make TEEs applicable in the HPC environment, crucial factors include minimal performance impact, compatibility with accelerators, and the absence of a requirement for application modifications. Dr. Peisert showed that certain commercial products exhibit negligible performance overhead. He also explored an open-source approach — Keystone TEE — for HPC systems. Attendees left with the conviction that data security will continue to advance with the help of these novel technologies.