HYPERION RESEARCH

HPC User Forum Update

Berkeley Lab and Collaborators Seek New Security Paradigm

Thomas Sorensen and Bob Sorensen
October 2021

IN THIS UPDATE

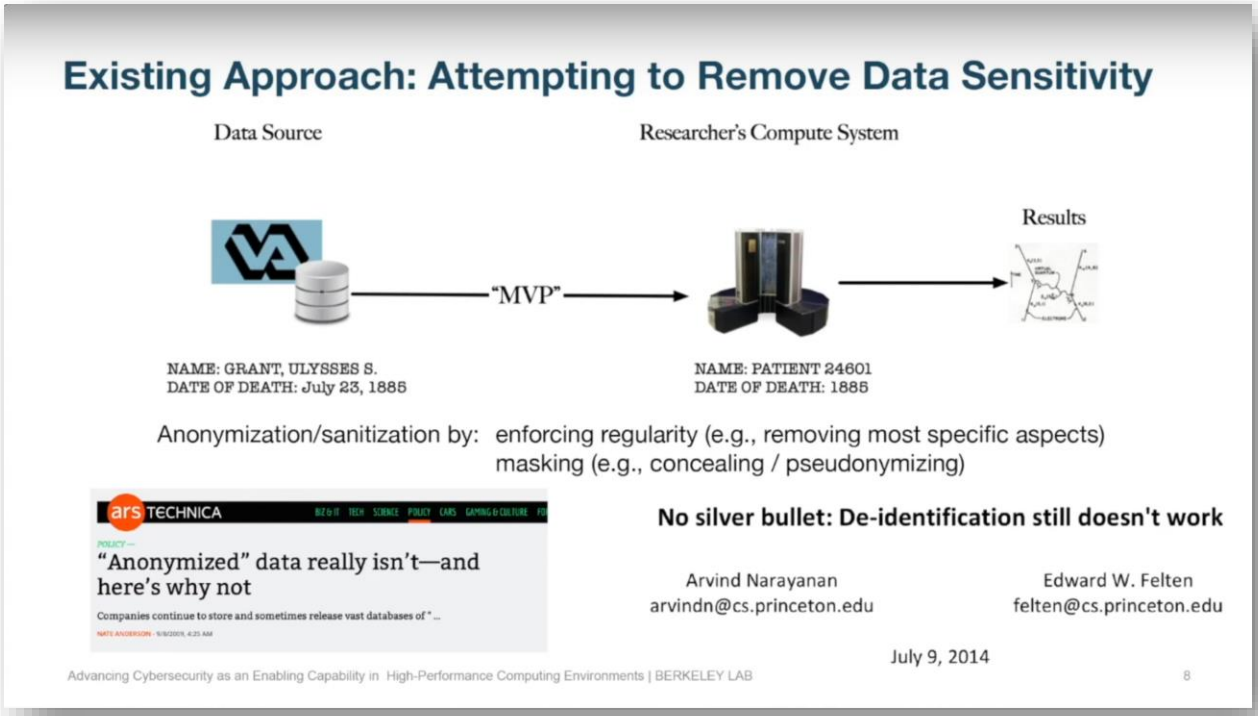
The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In September 2021, the 77th HPC User Forum took place virtually. This update summarizes a presentation from that virtual conference given by Dr. Sean Peisert, computer security research lead and staff scientist and Lawrence Berkeley National Laboratory. He provided insights on security in high-performance computing environments, specifically about how emerging technologies and research can advance cybersecurity as an enabling capability as opposed to an impedance to optimal performance or other barrier for users.

Source: Berkeley Lab 2021

PRESENTATION: BERKELEY LAB AND COLLABORATORS SEEK NEW SECURITY PARADIGM GIVEN BY DR. SEAN PEISERT, BERKELEY NATIONAL LAB

Peisert enumerated several of the many reasons that security in a scientific computing environment is valuable. First on his list was protecting data from the relatively new but now prevalent ransomware. This practice, which exploits the anonymity or near-anonymity of some crypto-currencies, is increasingly affecting high-performance computing environments. Other critical security priorities included protecting the integrity of data from modification or loss, protecting confidential data of any kind from exposure, and one that is of particular importance to high-performance systems, protection from misuse of compute resources. Theoretically, if a user were to use cycles meant for scientific research on something like mining cryptocurrency, that would represent not only a breach in security but potentially a significant loss of expensive compute resources.

FIGURE 1



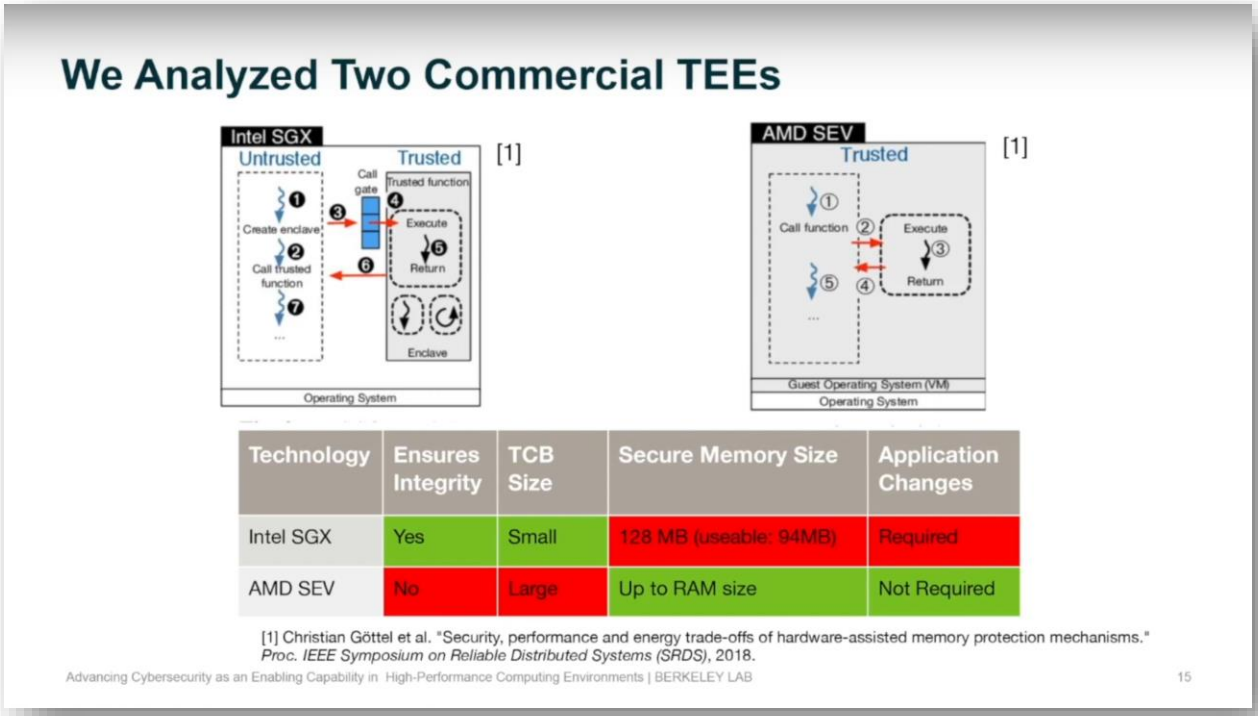
Source: Berkeley Lab 2021

For many, these threats are not just theoretical. In recent years, there have been incidents in the US, Russia, and Europe in which high-end systems have been misused for just this purpose. With new and unique methods emerging for breaching and exploiting systems and their data, Peisert looked to existing security models and broke down some of their weaknesses. As it is now, if one is going to offer data to an organization to do analysis in a supercomputing environment, there will either be a

trusted third party (this could be a public computing facility or a third-party vendor of some kind) or a legal agreement stipulating that the data will be protected in some way.

Citing numerous data breaches that have hit headlines in recent years, Peisert states that these common practices are not enough. There have been many cases of massive data breaches that occurred despite intent of a trusted party, legal obligations to protect, or some combination of the two. But for Peisert, the future of cybersecurity isn't simply addressing these issues, it is enabling greater capabilities and redefining data paradigms to keep up with the needs of high-performance applications. People and organizations are reluctant to implement confidential measures such as personally identifiable information or trade secrets, but this data, however, if properly anonymized and securely handled, can be a tremendous boon to research programs.

FIGURE 2



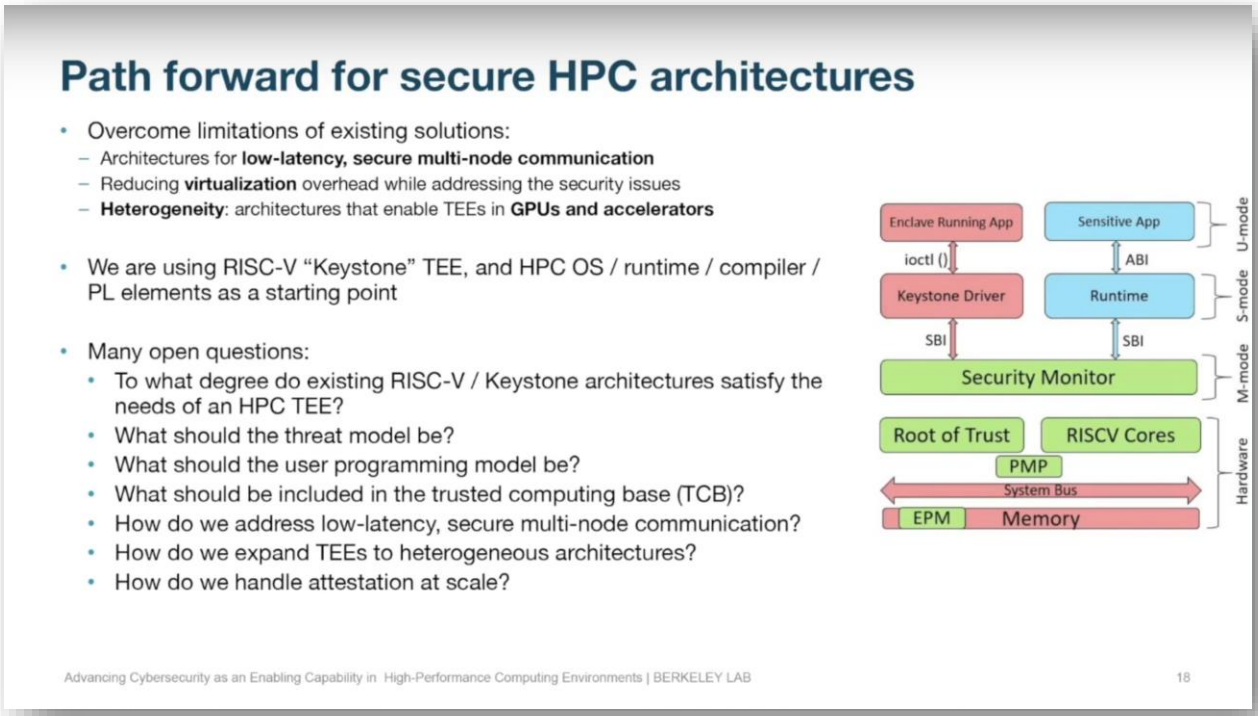
Source: Berkeley Lab 2021

So, Peisert asks, how does one compute with such a large amount of sensitive, regulated data? How can one protect against internal and external threats? How can the current trust model be reworked to enable compute research and better protect information? One current approach is to try to remove the sensitive aspects of the data. For instance, an identifiable name on a piece of healthcare data might be replaced by a unique number; an exact date may be replaced by just a month and year. This traditional method of sanitization has been shown not effective under scrutiny. According to Peisert, through inference attacks by linking external data sets or other information, anonymization really doesn't work.

Commonly, he states, either too much information is taken away, rendering it an insufficient research resource, or too much is left, allowing for personally identifiable information to be extracted.

Another existing approach, trust by physical protections, is when critical data is 'air gapped' or otherwise penned off from public networks and the internet. This method, while offering heightened protection, is often undesirable or untenable when conducting research with the intent of optimizing rapid scientific advancements. Most existing solutions focus on physical system isolation, which among other weaknesses, do not protect against internal threats. As it stands now, both HPC and cloud systems require full trust of the computing center and the system administrators to function securely.

FIGURE 3



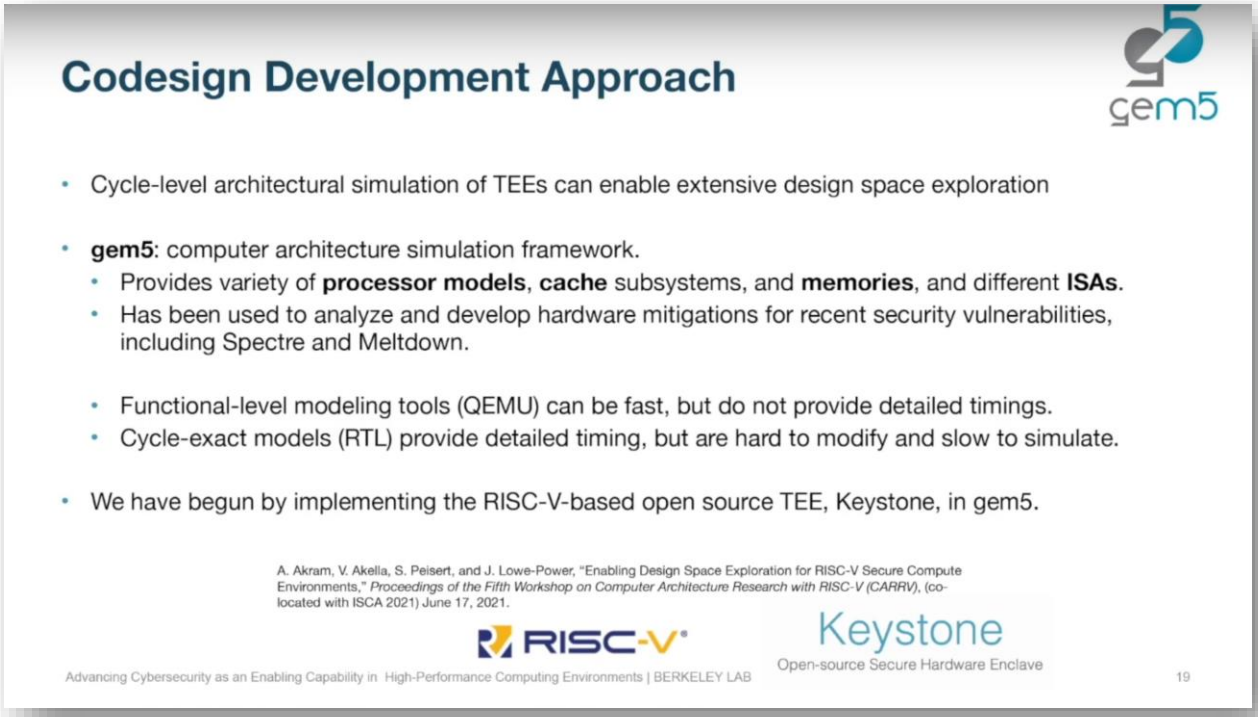
Source: Berkeley Lab 2021

To help solve this multi-faceted problem, Peisert looked to compare and contrast HPC security with that of traditional IT. They're both connected to IP networks and they both run Linux-like operating systems, but that is largely where the similarities end. In addition to the eponymous use of high-performance computation and data transfers, HPC systems tend to run on highly exotic hardware and software stacks. Additionally, the international and open nature of many collaborative HPC systems and environments would not allow for network firewalls, 'air gaps', and certain other security solutions.

Although these needs can present certain difficulties, features like exotic hardware and software stacks can be leveraged as an opportunity and be modified to add HPC-appropriate security. One element which Peisert and his team have been exploring is called Trusted Execution Environments (TEEs). TEEs enforce strong separation from other processes, can encrypt memory and computation, and can prevent exposure to other users and even sysadmins (system administrators). They enable a data owner to share data without having to trust the compute facility. Given the appropriate architecture, TEEs can enable strong security in HPC involving sensitive data. The entire threat model of a data provider having to trust the data center completely gets broken down with this. The TEEs, by encrypting memory and isolating from other users and even sysadmins, doesn't require that trust anymore. Trust executing exists and is already in use today.

CPU manufacturers like Intel, AMD, and Arm all make chips now that support some variation of TEEs (SGX, SEV, and Trustzone respectively). Additionally, there are several open source hardware implementations of TEEs, notably the RISC-V based Keystone environment from UC Berkeley. And, in what Peisert perceives as a testament to the power of this technology, all major cloud providers (including Google, Amazon, and Microsoft) implement confidential computing based on different TEE like technologies. There is also a Linux Foundation project called the Confidential Computing Consortium that is expanding the development of TEEs as well.

FIGURE 4



Source: Berkeley Lab and Hyperion Research, 2021

In an effort to investigate their potential efficacy in the world of HPC, Peisert and his team analyzed two commercial TEEs. They explored the Intel SGX and AMD SEV technologies with criteria such as integrity assurance, the size of the trusted computing bases, the secure memory size, and whether they require application changes. They took this process one step further by evaluating workloads in the form of running benchmarks.

Their results, after evaluating workloads of traditional nature as well as more modern jobs like graph analytics, indicated that Intel SGX was not appropriate for HPC. It slows down as much as two orders of magnitude compared with unsecure execution. This is a result of some low-level architectural elements as well as the fact that there is only 96MB of usable memory. AMD's SEV technology, however, proved performant in an HPC environment but was limited due to the virtualization overhead, lack of secure low-latency node communication, and that it only currently supports CPUs and no GPUs or accelerators.

Peisert and his collaborators are looking to develop new, secure architectures for HPC that overcome the limitations of existing solutions by addressing secure, low-latency multi-node communication, reducing virtualization overhead, and enabling heterogeneity. Their exploration and development process is based on the RISC-V Keystone TEE and are leveraging the Keystone runtime and compiler elements as a starting point and building on existing HPC OS, runtime, and compiler and programming language elements and trying to map them together. There are many questions to be answered, such as the ideal models for threat and user programming, what should be included in the trusted computing base, and how to handle attestation at scale.

For Peisert and his team, introducing security into these high-performance environments means more than protecting regulated or private data and compute. It is a means of enabling next-generation research through novel data management and security methods. Many of the solutions in traditional IT cannot be easily translated into these unique HPC architectures, indicating that new and performance-aware methods can help safely bring more valuable data to open, collaborative research projects. By taking data protection and privacy into a new and more secure place, efforts like these can help enable a greater number of advanced, data driven research projects.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
