LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LLSC News

Lincoln Laboratory Supercomputing Center

January 20, 2023

Top HPC Players Creating New Security Architecture Amid Neglect

Agam Shah | HPC Wire

Security of high-performance computers is being neglected in the pursuit of horsepower, and there are concerns that the ignorance may be costly if safeguards are not in place to protect these high-value systems.

The concerns are now being taken seriously in both the public and private sector, who are jointly defining a security architecture as part of a working group called High-Performance Computing Security, which is managed by National Institute of Standards and Technology and the National Science Foundation.

NIST organized the first-ever high-performance computing cybersecurity workshop at the Supercomputing 2022 conference (SC22) in November. The goal was to raise awareness and foster discussion on the topic, and to recruit more troops to arm supercomputers – some of which run applications to protect national interests – with better security measures.

"I have talked to a lot of people this week. When you say 'you know, how much do you care about security?' And it's like, 'absolutely none. I do not care about that. I do not want to know about it,'" said CJ Newburn, a distinguished engineer at Nvidia, during the workshop. He also reminded everyone that he was talking as a member of the HPC community, and his views did not represent those of Nvidia's positions on the matter.

Panelists and presenters at the workshop repeated common themes, which included the performance penalty of implementing security in systems. Many in the audience complained that system vendors were more interested in meeting performance benchmarks as stated in contracts, and that they ignored security as it would slow down systems.

"Performance and data security is a constant tussle between the vendors and the operators, and many vendors are reluctant to make these changes if the change negatively impacts the system performance," said Yang Guo, computer scientist at NIST and part of the HPCS working group.

The working group was established after executive order 13702 was passed in 2015, which established the National Strategy Computing Initiative to maximize the benefits of HPC for economic competitiveness and scientific discovery. The HPC security group was established in 2016 to create a security guidance and a management framework that provides a comprehensive and reliable security guidance to the HPC ecosystem.

The NIST working group is holding a meeting in March this year to push the security forward, and Guo is welcoming volunteers to join the committee. The goal is to get a diverse set of viewpoints from academia and the public and private sectors. Many labs with HPC systems are operating in silos, and one goal is to boost information exchange, Guo said.

Over the years, the system architecture has changed with the emergence of AI and newer modes of computing, which has led to revisions in the guidance. But the activity has picked up in recent years after a spate of cyberattacks on HPC systems.

One audience member at SC22 had a question as to whether any major U.S. HPC clusters had been compromised, which attracted laughter from the crowd as HPC installations typically do not reveal breaches. But one audience reminder gave an example of how a European supercomputer was hacked in 2020 to mine cryptocurrency.

Top HPC Players Creating New Security Architecture Amid Neglect (continued)

Just last week, AMD released a giant list of vulnerabilities in its first-, second- and third-generation Epyc chips, which have firmware vulnerabilities that can allow hackers to take control of systems by bypassing the security mechanisms in the boot sector. The world's fastest-ranked supercomputer, called Frontier, runs on AMD's third-generation Epyc chips, but it is not clear if the supercomputer was affected.

Implementing security like antivirus on PCs isn't the same as HPC, Guo said. That is because the attributes of HPC, such as large storage size and system access, are very different from conventional computing.

HPC security measures include "how to support the container and how to secure them, how to sanitize the computer nodes that pertain to projects on HPC so on and so forth," Guo said. He added that HPC was a shared resource, and high availability of systems and horsepower are equally or more important.

The NIST working group is looking at the HPC architecture and conducting risk assessment. A special NIST publication being written will provide a baseline and lexicon for HPC security. "We are trying to wrap up in the near future," Guo said.

The HPC reference model is a modified version of security measures adopted by MIT's Lincoln Laboratory, a United States Department of Defense federally-funded center. The entire system is broken up into four function zones, which include the access zone, management zone, high performance computing zone, and data storage. Different function zones provide different services and face different threats, and nodes in each function zone are more uniform, Guo said.

"The software around them is different from one to the other, which means that we potentially can have different security guidance for different zones, which will alleviate the burden of applying one security guidance for all the nodes within a complex HPC system," Guo said.

The architecture guidance is a great start for auditors and governance staff to gain an understanding of a baseline HPC system. It will form an overlay to create an overarching security plan that would cover the entire system, "but since the controls can be identifiers, we can specify in what zone or in what new type of your system it is applicable," said Catherine Hinton, cybersecurity lead for High Performance Computing at Los Alamos National Laboratory.

Panel members also talked about how they were working to boost cybersecurity in systems.

MIT initially limits access to HPC hardware to new users and observes them over time before giving them access to more resources. On the campus system, an initial user only gets access to about two nodes, or 96 cores in total, and their requests are constantly evaluated and vetted.

"They then get to run on another slightly larger allocation, but still not especially large as they're working with their code to parallelize it. And we get to see what they're doing, to some degree," said MIT's Albert Reuther, a senior staff member at the Lincoln supercomputing center.

Over time, users get access to more resources. There is always a second person vetting requests to access the supercomputers, like where they are from and whether there are legitimate requests, Reuther said, adding "but it doesn't completely eliminate the risk."

MIT, which has 11 HPC engineers working across three clusters, also got rid of root access to systems. Instead, the system administrators have root privileges through "sudo," a shell command that provides a way to audit everything engineers do on the system. Sudo records everything and leaves a paper trail that can help track down abnormal behavior.

The U.S. Department of Defense's HPCMP (High Performance Computing Modernization Program) has five security centers for its five different supercomputing centers. The defense arms, including Navy, Air Force and Army, have their own security rules, said Rickey Gregg, HPCMP's cybersecurity program manager.

"I tried to make it as simple as possible across all of these centers to have a common lexicon and common … controls that they need to apply to make it as efficient as possible. I know security is often the bad guy in the room, but we really try to help as much as we can," Gregg said.

Amazon suggested another way to integrate security as a broader set of plugins and microservices, especially with HPC moving into the cloud. Modern approaches to HPC security are also needed as the boundaries of computing expand into edge devices and cloud services.

"We often try to think of HPC systems as being a walled garden. You set up a barrier, your front ends, your access points, etc. And everything behind that gets to be the Wild West," said Lowell Wofford, principal specialist solution architect for High Performance Computing at AWS, at the SC workshop.

AWS recently introduced a new EC2 offering called Hpc7g instances, which is for high-performance computing in the cloud. The offering is based on Amazon's first HPC CPUs called Graviton3E, and the instance also runs in a controller called Nitro V5, which has in-built security that the company claims does not slow down performance.

Nitro's root of trust secures data in transit and as it is being processed in the AWS instance. The security chip protects firmware by ensuring its hash representation is valid, and uses trusted keys and encryption to prevent tampering with the data. A security API for cloud-based microservices can also be attached to HPC applications.

AWS's approach is to migrate security from its commercial cloud service to the HPC offerings, Wofford said.

"We have done so by applying the same methods we always use, but enhancing our services for tight integration. And one of the results of that is we can continue to apply the security patterns that we have always used in the cloud as we migrate into HPC," Wofford said.

Nvidia's answer to security is to do what HPC does well: accelerate. The company has security tools like its Morpheus AI toolkit that can identify uneven patterns in computing systems, which it can then highlight to security teams. The Morpheus toolkit relies on Nvidia GPU acceleration.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.