Contribution ID: **7** Type: **Talk**

Introducing Cloud-Native Supercomputing: Bare-Metal, Secured Supercomputing Architecture

Thursday, 2 December 2021 11:00 (30 minutes)

High-performance computing and artificial intelligence have driven supercomputers into wide commercial use as the primary data processing engines enabling research, scientific discoveries, and product development. Extracting the highest possible performance from supercomputing systems while achieving efficient utilization has traditionally been incompatible with the secured, multi-tenant architecture of modern cloud computing. A cloud-native supercomputing platform aims at the goal of combining peak system performance with a modern zero-trust model for security isolation and multi-tenancy. The key element enabling this architecture transition is the data processing unit (DPU).

The DPU is a fully integrated data-center-on-a-chip platform that imbues each supercomputing node with two new capabilities: First, an infrastructure control plane
processor that secures user access, storage access, networking, and life-cycle orchestration for the computing node in the data center or at the edge, offloading
these services from the main compute processor and enabling bare-metal multi-tenancy. Second, an isolated line-rate data path with hardware acceleration that
enables high performance. All this infrastructure allows a cloud-native HPC and
AI platform architecture that delivers HPC performance on an infrastructure
platform that meets cloud services requirements. The implementation of the infrastructure comes from the open-source community and driven by standards, similarly as how some of the traditional HPC software stack that is maintained by a community including commercial companies, academic organizations, and government agencies.

We'll introduce the new supercomputing architecture, discuss the first cloud native supercomputers, , review first applications performance results, and explore future directions.

Student?

No

Supervisor name

Supervisor email

Primary author: Mr SHAINER, Gilad (NVIDIA)

Co-author: SLAMA, David

Presenters: Mr SHAINER, Gilad (NVIDIA); SLAMA, David

Session Classification: HPC Technology

Track Classification: HPC Technology