Keynote 1

The Data Center is a Computer

Michael Kagan, Chief Technology Officer, *NVIDIA*

Abstract:

High performance computing and Artificial Intelligence are the most essential tools fueling the advancement of science. NVIDIA GPU and Networking technologies are the engines of the modern HPC data center, delivering breakthrough performance and scalability. In order to handle the ever growing demands for higher computation performance and the increase in the complexity of scientific problems, a new data processing unit (DPU) was created. DPUs are interconnect elements that include In-Network Computing engines, engines that can participate in the application run time by analyzing data on the fly. The combination of CPUs, GPUs, and DPUs, creates the next generation of data center and edge computing architectures.

Bio:

Michael Kagan is a co-founder of Mellanox and has served as CTO since January 2009. Previously, Mr. Kagan served as vice president of architecture from May 1999 to December 2008. From August 1983 to April 1999, Mr. Kagan held a number of architecture and design positions at Intel Corporation. While at Intel Corporation, between March 1993 and June 1996, Mr. Kagan managed Pentium MMX design, and from July 1996 to April 1999, he managed the architecture team of the Basic PC product group. Mr. Kagan holds a Bachelor of Science in Electrical Engineering from the Technion - Israel Institute of Technology.

Keynote 2

Scalable Reliable Datagrams: Combining High Performance Computing and Cloud-Scale Networks

Brian Barrett, Principal Engineer, Networking, *AWS*

Abstract:

Traditionally, Cloud providers have built extremely large Ethernet networks focused on good overall performance for web and enterprise applications. Scale and availability for web services-type applications are primary design criteria. High Performance Computing networks, such as InfiniBand, focus on high throughput and low latency, with scaling and availability as secondary concerns. Traditionally, Cloud providers have relied on their scalable Ethernet networks and TCP to address their customers' HPC needs. While this works well for small-scale applications, its shortcomings quickly become obvious. More recently, Cloud providers have provided small segments of their offering with customized InfiniBand offerings to provide the peak performance, but giving up some of their traditional scaling capabilities.

Amazon Web Services took a different approach to addressing HPC requirements without building smaller, custom HPC networks. The result is Elastic Fabric Adapter (EFA) and a new AWS-specific reliability protocol called Scalable Reliable Datagrams (SRD). SRD allows AWS to provide strong HPC application scalability across multiple hardware types on a single network. Building EFA and SRD required not only a number of hardware challenges, but also working through a number of challenges in the HPC software ecosystem. While EFA is still a relatively new network device, we are already seeing positive results with a number of common HPC applications.

Bio:

Brian Barrett is a Principal Engineer in the High Performance Computing organization at Amazon Web Services (AWS). While at AWS, Brian lead the development of the Elastic Fabric Adapter as well as the networking stack for the Nitro hypervisor system. Prior to joining AWS, Brian was a Principal Member of Technical Staff at Sandia National Laboratories, where his networking research included work with both the Portals 4 Network Programming Interface and the Open MPI implementation of the Message Passing Interface. Brian received his Ph.D. in Computer Science from Indiana University, Bloomington in 2009.