CASE STUDY

Reaching the Summit with InfiniBand:

Mellanox Interconnect Accelerates World's Fastest HPC and Artificial Intelligence Supercomputer at Oak Ridge National Laboratory (ORNL)

ABOUT SUMMIT

On June 8, 2018, the U.S. Department of Energy's Oak Ridge National Laboratory (ORNL) launched Summit as "the world's most powerful and smartest scientific supercomputer." Clocked at 200 petaflops, Summit provides eight times higher processing speed than its predecessor, Titan, the previous top-ranking, fastest supercomputer in the US. Summit enables higher resolution and fidelity simulations that advance human knowledge in diverse science domains such as biology, nuclear science, cosmology and more. Initial Summit projects will include the modeling of exploding stars at unprecedented scales, simulating particle turbulence in sustainable fusion reactions, conducting research into materials for high-temperature superconductors, and more.

As the fastest artificial intelligence (AI) platform in the world, Summit offers unparalleled opportunities to integrate AI with scientific discovery. For scientists, the ability to apply machine learning and deep learning capabilities to automate, accelerate, and drive understanding at supercomputer-scale is especially critical as it facilitates breakthroughs in human health, energy and engineering, while also answering fundamental questions about the universe.

https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

Figure 1. Summit Supercomputer (ORNL)

HIGHLIGHTS

- Collaboration between ORNL, IBM, Mellanox, NVIDIA and Red Hat
- 200 petaflops – world's highest processing speed
- World's fastest artificial intelligence platform
- Mellanox end-to-end dual-rail EDR 100Gb/s InfiniBand solution including ConnectX®-5 adapter cards, Switch-IB™ 2-based switch systems, LinkX® cables and modules, software, and services.
- 10x higher performance: In-Network Computing technology runs algorithms on data as it moves through the network
- InfiniBand 'smart accelerations' & offload technology, delivering the highest HPC and AI application performance, scalability, and robustness
- End-to-end Interconnect deployment delivery by Mellanox Professional Services

"Summit HPC and AI-optimized infrastructure enables us to analyze massive amounts of data to better understand world phenomena, to enable new discoveries and to create advanced AI software. InfiniBand In-Network Computing technology is a critical new technology that helps Summit achieve our scientific and research goals.

We are excited to see the fruits of our collaboration with Mellanox over the last several years through the development of the In-Network Computing technology, and look forward to take advantage of it for achieving highest performance and efficiency for our applications."

– Buddy Bland, Program Director at Oak Ridge Leadership Computing Faculty

CHALLENGING OLD ARCHITECTURAL APPROACHES

The 2014 requirements of the CORAL program – the US Department of Energy's Collaboration of Oak Ridge, Argonne, and Lawrence Livermore labs – necessitated a new approach to computing. By then, it was already a well-established fact that the traditional CPU-centric/powerful server approach to computing could not satisfy the demand to analyze a growing amount of data, support complex simulations, overcome performance bottlenecks, and create intelligent data algorithms. Next-generation performance goals have driven the need for a new, data-centric approach – a system solution that furnishes compute power wherever data resides in the system. Such an approach enables the system to manage and carry out computational operations on the data as it is being transferred between the system's various compute, storage, and interconnect elements.

The opportunity to implement CORAL's Summit supercomputer was awarded to IBM, NVIDIA and Mellanox, whose joint solution demonstrates a data-centric system approach that enables convergence of infrastructure analytics, modeling, visualization and simulation, unleashing new scales of speed and paths to innovation.

SUMMIT ARCHITECTURE

Summit's hybrid CPU-GPU architecture consists of 4,608 compute servers (nodes), each containing two 22-core IBM Power9 processors and six NVIDIA Tesla V100 graphics processing unit accelerators; these are interconnected with dual-rail Mellanox EDR 100Gb/s InfiniBand technology. Summit also houses more than 10 petabytes of memory with fast, high-bandwidth pathways that allow moving data efficiently.

Each node has over half a terabyte of coherent memory (high bandwidth memory + DDR4) addressable by all CPUs and GPUs, plus 800GB of non-volatile RAM that can be used as a burst buffer or as extended memory. To provide a high rate of data throughput, the nodes are connected in a non-blocking fat-tree topology using dual-rail Mellanox EDR 100Gb/s InfiniBand interconnect for both storage and inter-process communications traffic, delivering 200Gb/s bandwidth between nodes and also In-Network Computing acceleration for communications frameworks such as MPI and SHMEM/PGAS.

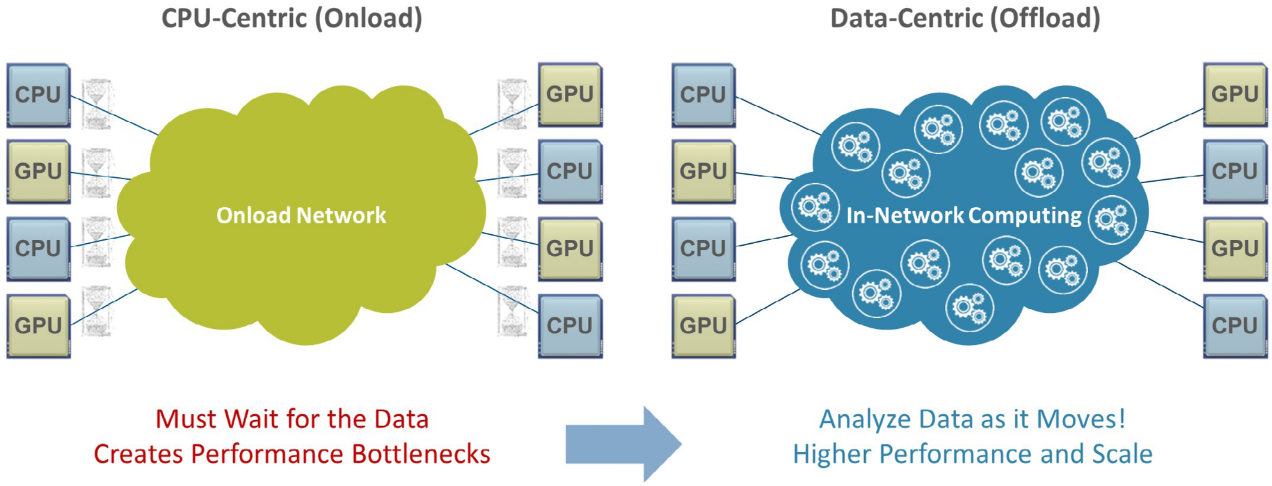

Figure 2. Moving from CPU-centric computing to data-centric computing.

MELLANOX EDR INFINIBAND SOLUTION

The need to analyze growing amounts of data in order to support complex simulations, overcome performance bottlenecks and create intelligent data algorithms requires the ability to manage and carry out computational operations on the data as it is being transferred by the data center interconnect. Mellanox's end-to-end EDR 100Gb/s InfiniBand solution provides the interconnect for Summit's storage and inter-process communications traffic. Incorporating In-Network Computing technology that performs data algorithms within the network devices, it delivers ten times higher performance than CPU-centric computing, while enabling the era of "data-centric" data centers.

"We are proud to accelerate the world's top HPC and AI supercomputer at the Oak Ridge National Laboratory, a result of a great collaboration over the last few years between Oak Ridge National Laboratory, IBM, NVIDIA and us. Our InfiniBand smart accelerations and offload technology deliver the highest HPC and AI applications performance, scalability, and robustness.

InfiniBand enables organizations to maximize their data center return-on-investment and improve their total cost of ownership and, as such, it connects many of the top HPC and AI infrastructures around the world. We look forward to drive and to accelerate new scientific discoveries and advances in AI development to be empowered by Summit."

– Eyal Waldman, President and CEO of Mellanox Technologies

Key Features of the Mellanox EDR InfiniBand Solution:

Adaptive Routing

The scalability improvements in Mellanox EDR 100Gb/s InfiniBand include support for Adaptive Routing for highest network efficiency.

SHIELD (Self Healing Interconnect Enhancement for inteLligent Datacenters)

SHIELD enables switches to exchange real-time information to overcome link failures and optimize data flows, with absolutely no performance overhead. SHIELD enables recovery from link failures 5000x faster than a traditional software-based subnet manager could recalculate and distribute an adjusted routing table.

https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

Figure 3: Summit houses more than 10 petabytes of memory with fast, high-bandwidth pathways that allow moving data efficiently.

NVMe Burst Buffer

ConnectX-5 provides offloaded access to NVMe target memory. This supports offloading the movement of data to / from the NVMe device without the need for CPU intervention. This capability also supports background data staging.

Mellanox SHARP ™

Mellanox Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ technology improves upon the performance of MPI operations by offloading collective operations from the CPU to the switch network, and by eliminating the need to send data multiple times between endpoints. This innovative approach decreases the amount of data traversing the network as aggregation nodes are reached, and dramatically reduces the MPI collective operations time. Implementing collective communication algorithms in the network also has additional benefits, such as freeing up valuable CPU resources for computation rather than using them to process communication.

Professional Services - Summit

The end-to-end deployment, acceptance testing and performance tuning by the Mellanox Professional Services team contributed to the success of the Summit project. Mellanox Professional Services assumed end-to-end ownership and responsibility of the project network infrastructure, starting with the point-to-point cable plans and off-site cable labeling and bundling, through the on-site physical installation and logical configuration, and on to the official interconnect site-certification, with experts supporting both performance and acceptance tests, using an advanced performance tuning platform tailored to the project's needs.

Facing a myriad of on-site challenges, a dynamic environment, new technologies, and a tight schedule—the collaborative effort between the Mellanox Professional Services team and IBM Lab Services team resulted in making the project an on-time success.

Table 1 - Summit Features*

Summit Feature	Description
Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs; 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4

*https://www.olcf.ornl.gov/wp-content/uploads/2018/06/Summit_bythenumbers_FIN-1.pdf

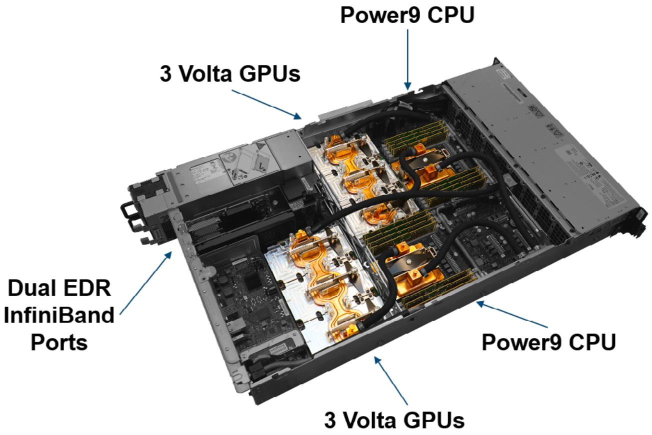

Figure 4. Summit Server Configuration

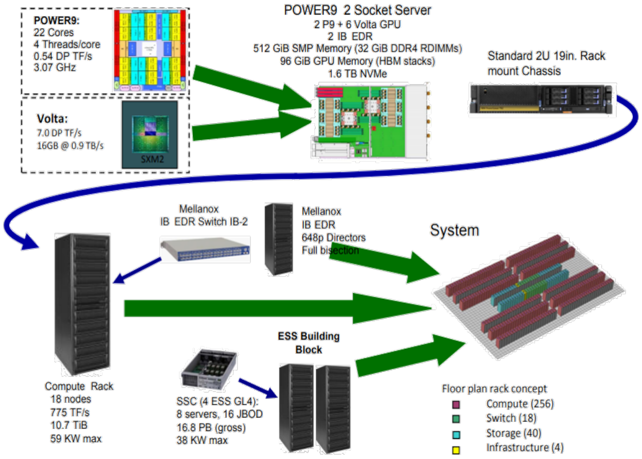

Figure 5. Summit High-Level Overview

About Mellanox

Mellanox Technologies (NASDAQ: MLNX) is a leading supplier of end-to-end InfiniBand and Ethernet smart interconnect solutions and services for servers and storage. Mellanox offers a choice of fast interconnect products: adapters, switches, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services.
More information is available at www.mellanox.com

About Oak Ridge National Laboratory (ORNL)

Oak Ridge National Laboratory is managed by UT-Battelle for the Department of Energy's Office of Science, the single largest supporter of basic research in the physical sciences in the United States. DOE's Office of Science is working to address some of the most pressing challenges of our time. For more information on Summit, please check:
https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com