

How the Results of Summit and Sierra are Influencing Exascale

AI Geist

Oak Ridge National Laboratory

HPC Forum

September 9-11, 2019

ORNL is managed by UT-Battelle
for the US Department of Energy

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.



ORNL Summit: IBM System Overview

System Performance

- Peak of 200 Petaflops (FP_{64}) for modeling & simulation
- Peak of 3.3 ExaOps (FP_{16}) for data analytics and artificial intelligence

The system includes

- 4,608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM file system transferring data at 2.5 TB/s

Each node has

- 2 IBM POWER9 processors
- 6 NVIDIA Tesla V100 GPUs
- 608 GB of fast memory (96 GB HBM2 + 512 GB DDR4)
- 1.6 TB of non-volatile memory



LLNL Sierra: IBM System Overview)

(Same parts different architecture)

System Performance

- Peak of 125 Petaflops (FP_{64}) for modeling & simulation
- Peak of 2.0 ExaOps (FP_{16}) for data analytics and artificial intelligence

The system includes

- 4,320 nodes
- Single plane Mellanox EDR InfiniBand network w/ 2 to 1 tapered Fat Tree
- 154 PB IBM file system transferring data at 1.5 TB/s

Each node has

- 2 IBM POWER9 processors
- 4 NVIDIA Tesla V100 GPUs
- 576 GB of fast memory (64 GB HBM2 + 512 GB DDR4)
- 1.6 TB of non-volatile memory



Summit Node

- 3 GPU per CPU
- Coherent memory across entire node
- 1.6 TB of on-node NVM

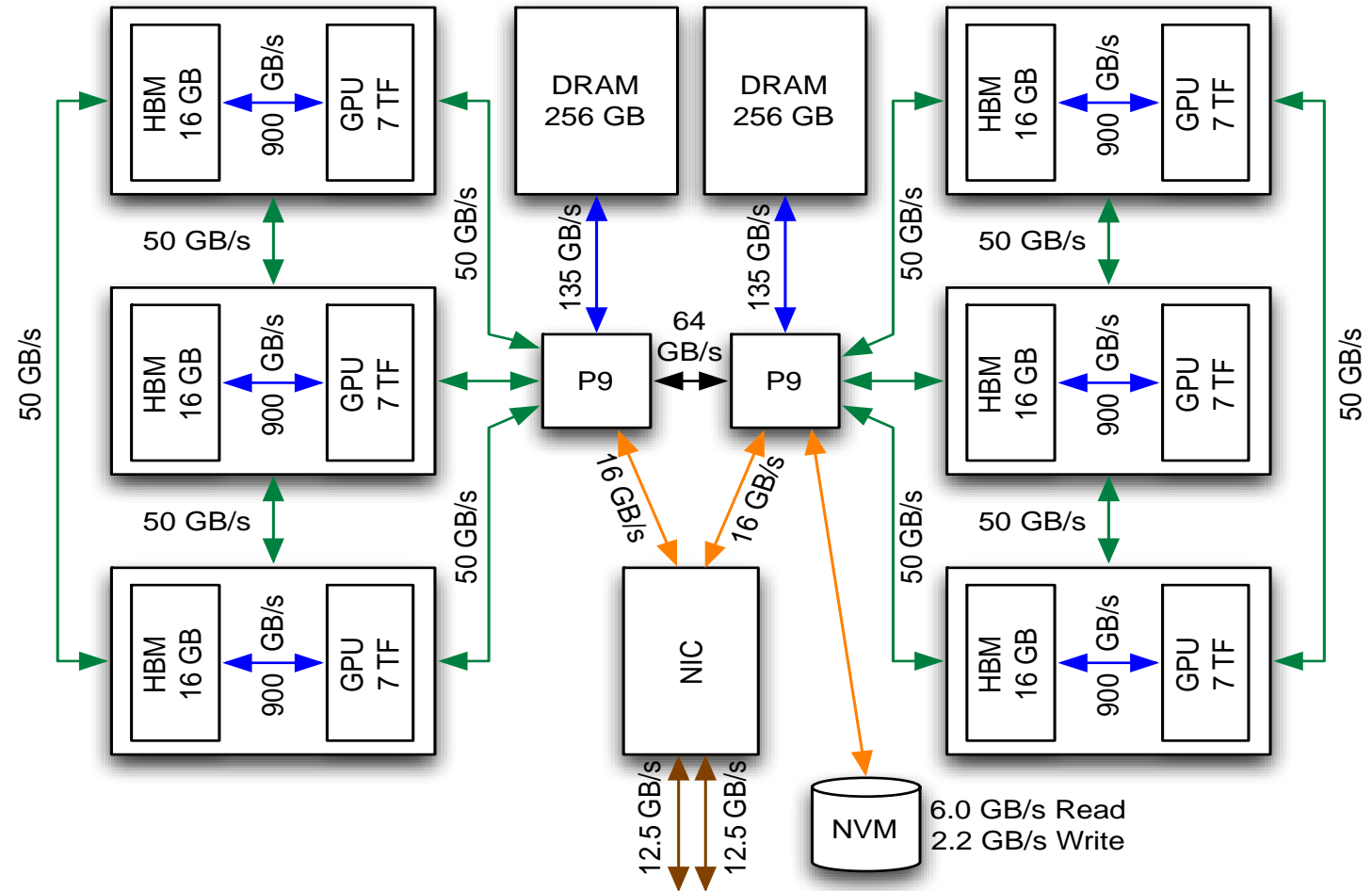
Sierra Node

- 2 GPU per CPU
- 75 GB/s GPU-CPU links

Titan Node

- 1 GPU per CPU
- AMD CPU / Nvidia GPU

6 GPU, 2 CPU Summit Node



TF
HBM
DRAM
NET
MMsg/s

42 TF (6x7 TF)
96 GB (6x16 GB)
512 GB (2x16x16 GB)
25 GB/s (2x12.5 GB/s)
83

HBM/DRAM Bus (aggregate B/W)
 NVLINK
 X-Bus (SMP)
 PCIe Gen4
 EDR IB

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

Summit has 27,648 NVIDIA Volta GPUs each with optimized AI Performance

Each Volta GPU can perform:

- 7.5 FP₆₄ TFLOPS | 15 FP₃₂ TFLOPS | 120 FP₁₆ TFLOPS
- Tensor cores do mixed precision multiply-add of 4x4 matrices



$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 FP16 FP16 FP16 or FP32

$$D = AB + C$$

- The Modeling & Simulation community can benefit by utilizing mixed / reduced precision algorithms
- AI community can do ML training at 120 FP₁₆ TFLOPs

Supercomputer Specialization vs ORNL Summit

- As supercomputers got larger and larger, we expected them to be more specialized and limited to just a small number of applications that can exploit their growing scale
- Summit's architecture with powerful, multiple-GPU nodes with huge memory per node seems to have stumbled into a design that has broad capability across:
 - Traditional HPC modeling and simulation
 - High performance data analytics
 - Artificial Intelligence

Summit Excels Across Simulation, Analytics, AI



- Data analytics – CoMet bioinformatics application for comparative genomics. Used to find sets of genes that are related to a trait or disease in a population. Exploits cuBLAS and Volta tensor cores to **solve this problem 5 orders of magnitude faster than previous state-of-art code.**
 - **Has achieved 2.36 ExaOps** mixed precision (FP₁₆-FP₃₂) on Summit
- Deep Learning – global climate simulations use a half-precision version of the DeepLabv3+ neural network to learn to detect extreme weather patterns in the output
 - **Has achieved a sustained throughput of 1.0 ExaOps (FP₁₆)** on Summit
- Nonlinear dynamic low-order unstructured finite-element solver accelerated using mixed precision (FP₁₆ thru FP₆₄) and AI generated preconditioner. Answer in FP₆₄
 - **Has achieved 25.3 fold speedup** on Japan earthquake – city structures simulation
- **Half-dozen apps >25x speedup on Summit vs. Titan (a couple around 100x speedup)**

Summit Displays Its Balanced Design Achieves #1 on TOP500, #1 on HPCG, and #1 Green500 (level 3)



122 PF HPL
Shows DP performance



2.9 PF HPCG
Shows fast data movement



13.889 GF/W
Shows energy efficiency

What Makes Summit Architecture Better Than Titan?

(These same concepts are being carried over into Frontier)



- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and higher memory bandwidth
- Faster interconnect
- Much higher bandwidth between CPUs and GPUs
- Much larger and faster file system
- 7x more performance for only slightly more power (HPL 122 PF run was 8.8 MW)

Feature	Titan	Summit
Peak FLOPS ₆₄	27 PF	200 PF
Max possible Power	9 MW	13 MW
Number of Nodes	18,688	4,608
Node performance	1.4 TF	42 TF
Memory per Node	32 GB DDR3 6 GB GDDR5	512 GB DDR4 96 GB HBM2
NV memory per Node	0	1.6 TB
Total System Memory	0.7 PB	2.8 PB + 7.4 PB NVM
System Interconnect	Gemini (6.4 GB/s)	Dual Rail EDR (25 GB/s)
Interconnect Topology	3D Torus	Non-blocking Fat Tree
Bi-Section Bandwidth	15.6 TB/s	115.2 TB/s
Processors on node	1 AMD Opteron™ 1 NVIDIA Kepler™	2 IBM POWER9™ 6 NVIDIA Volta™
File System	32 PB, 1 TB/s, Lustre®	250 PB, 2.5 TB/s, GPFS™

Frontier Continues the Accelerated Node Design begun at ORNL with Titan and continued with Summit

Frontier Node -- 4 GPUs per CPU:

- One purpose-built AMD EPYC™ processor
- Four HPC and AI optimized AMD Radeon Instinct™ GPUs
- Fully connected with high speed AMD Infinity Fabric links
- Coherent memory across the node
- 100 GB/s node injection bandwidth
- On-node NVM storage



Partnership between ORNL, Cray, and AMD

The Frontier system will be delivered in 2021

Peak FP_{64} Performance greater than 1.5 EF

Max Power Consumption 29 MW

Cray Shasta cabinets Connected by Slingshot™ interconnect

- with adaptive routing, congestion control, and quality of service

Questions?

ORNL / Cray / AMD Partnership

