Summit Architecture Overview

Scott Atchley

Oak Ridge National Laboratory

Summit Training Workshop 2018

Knoxville TN

February 11, 2019

ORNL is managed by UT-Battelle, LLC 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 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

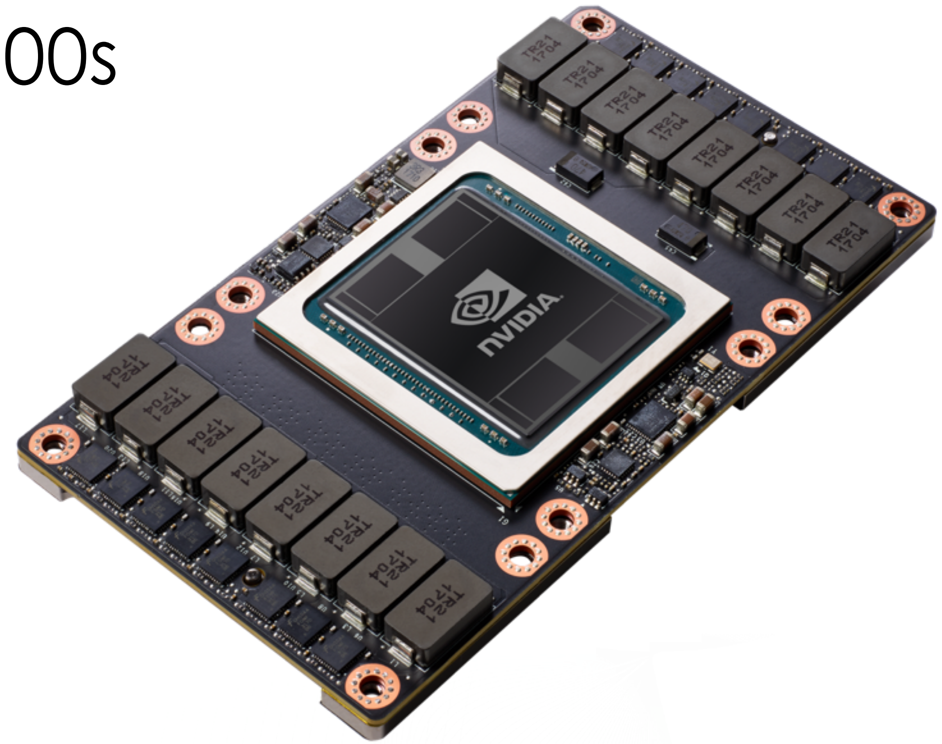
The system includes

- 4,608 nodes
- Dual-port 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 NV memory

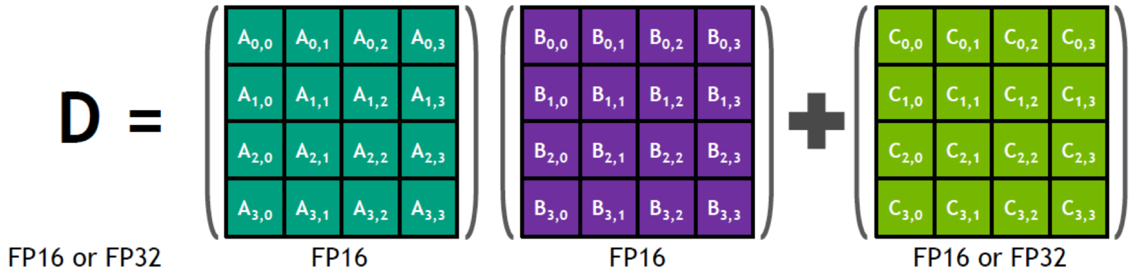
Summit Contains 27,648 NVIDIA Tesla v100s



Each Tesla v100 GPU has:

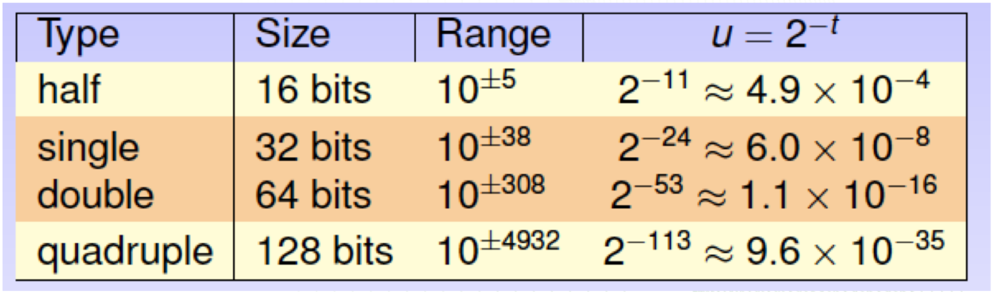
- 150+150 GB/s total BW (NVLink v2.0)
- 5,120 CUDA cores (64 on each of 80 SMs)
- 640 Tensor cores (8 on each of 80 SMs)
- 20MB Registers | 16MB Cache | 16GB HBM2 @ 900 GB/s
- 7.5 DP TFLOPS | 15 SP TFLOPS | 120 FP_{16} TFLOPS

- Tensor cores do mixed precision multiply-add of 4x4



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

$$D = AB + C$$



Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- The M&S community must figure how out to better utilize mixed / reduced precisions
- Eg: Possible to achieve 4x FP64 peak for 64bit LU on V100 with iterative mixed precision (Dongarra et al.)

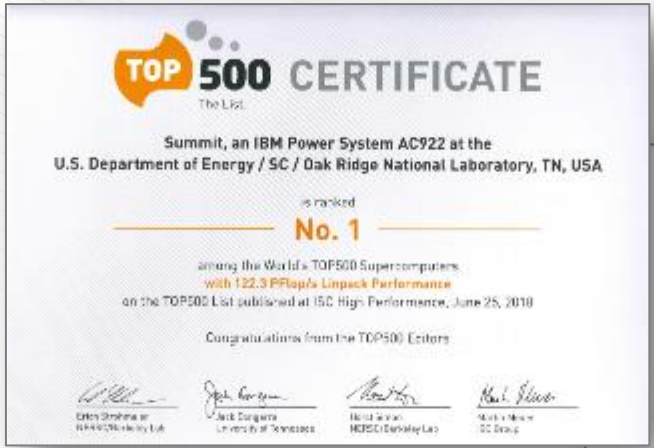
Supercomputer Specialization vs ORNL Summit

- 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 seems to have stumbled into a sweet spot that has broad capability across:
  - Traditional HPC modeling and simulation
  - High performance data analytics
  - Artificial Intelligence

In 2018 Summit Demonstrated Its Balanced Design

Achieves #1 on TOP500, #1 on HPCG, #1 Green500, and #1 on I/O 500



TOP 500 CERTIFICATE

The List

Summit, an IBM Power System AC922 at the
U.S. Department of Energy / SC / Oak Ridge National Laboratory, TN, USA

is ranked

No. 1

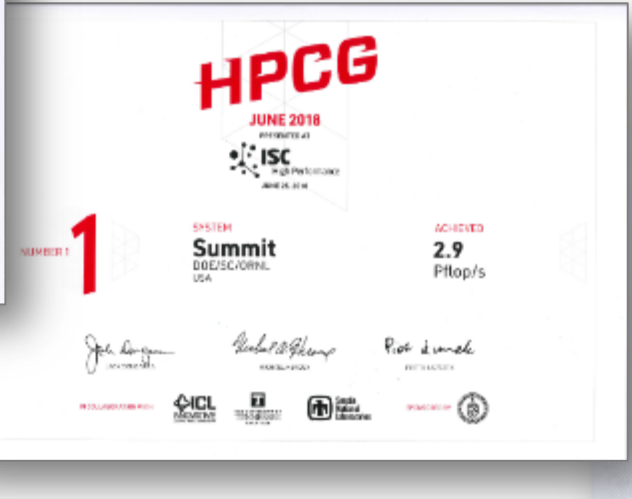
among the World's TOP500 Supercomputers
with 122.3 PFlop/s Linpack Performance
on the TOP500 List published at ISC High Performance, June 25, 2018

Congratulations from the TOP500 Editors

Erich Strohmaier, NERSC/Berkeley Lab — Jack Dongarra, University of Tennessee — Horst Simon, NERSC/Berkeley Lab — Martin Meuer, ISC Group

122 PF HPL
#1 raw performance

144 PF in Nov 2018



HPCG

JUNE 2018

ISC High Performance

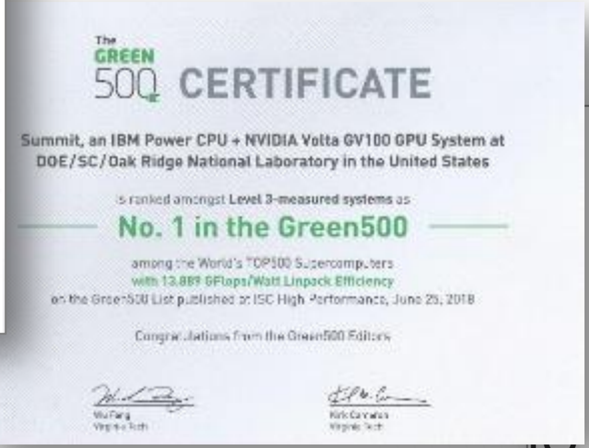
NUMBER 1

SYSTEM
Summit
DOE/SC/ORNL
USA

ACHIEVED
2.9
Pflop/s

ICL

2.9 PF HPCG
#1 fast data movement



The GREEN 500 CERTIFICATE

Summit, an IBM Power CPU + NVIDIA Volta GV100 GPU System at
DOE/SC/Oak Ridge National Laboratory in the United States

is ranked amongst Level 3-measured systems as

No. 1 in the Green500

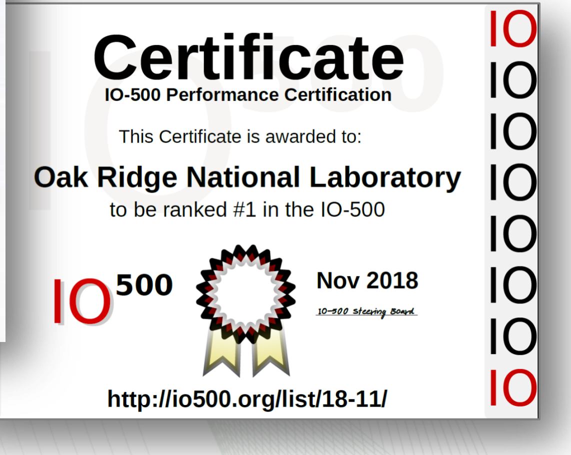
among the World's TOP500 Supercomputers
with 13.889 GFlops/Watt Linpack Efficiency
on the Green500 List published at ISC High Performance, June 25, 2018

Congratulations from the Green500 Editors

Wu Feng, Virginia Tech — Kirk Cameron, Virginia Tech

13.889 GF/W
#1 energy efficiency

14.668 GF/W Nov 2018



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

Oak Ridge National Laboratory

to be ranked #1 in the IO-500

IO500

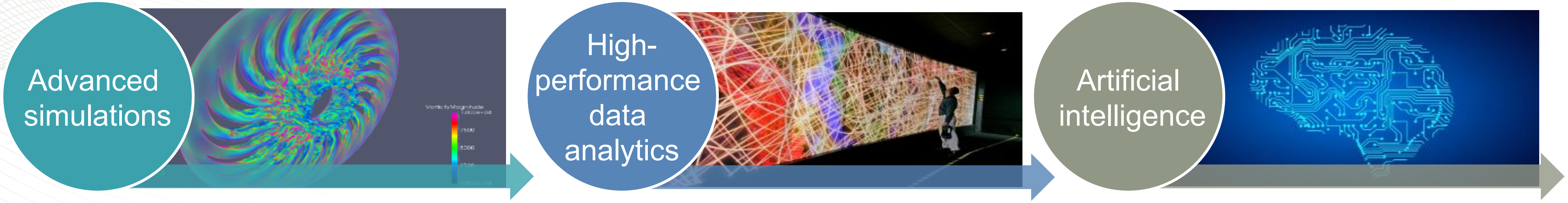
Nov 2018

IO-500 Steering Board

http://io500.org/list/18-11/

#1 HPC storage performance

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 (FP_{16}-FP_{32}) on Summit
- Deep Learning – global climate simulations use a half-precision version of the DeepLabv3+ neural network to learn to detecting extreme weather patterns in the output
  - **Has achieved a sustained throughput of 1.0 ExaOps (FP_{16})** on Summit
- Nonlinear dynamic low-order unstructured finite-element solver accelerated using mixed precision (FP_{16} thru FP_{64}) and AI generated preconditioner. Answer in FP_{64}
  - **Has achieved 25.3 fold speedup** on Japan earthquake – city structures simulation
- **Half-dozen Early Science codes are reporting >25x speedup on Summit vs Titan**

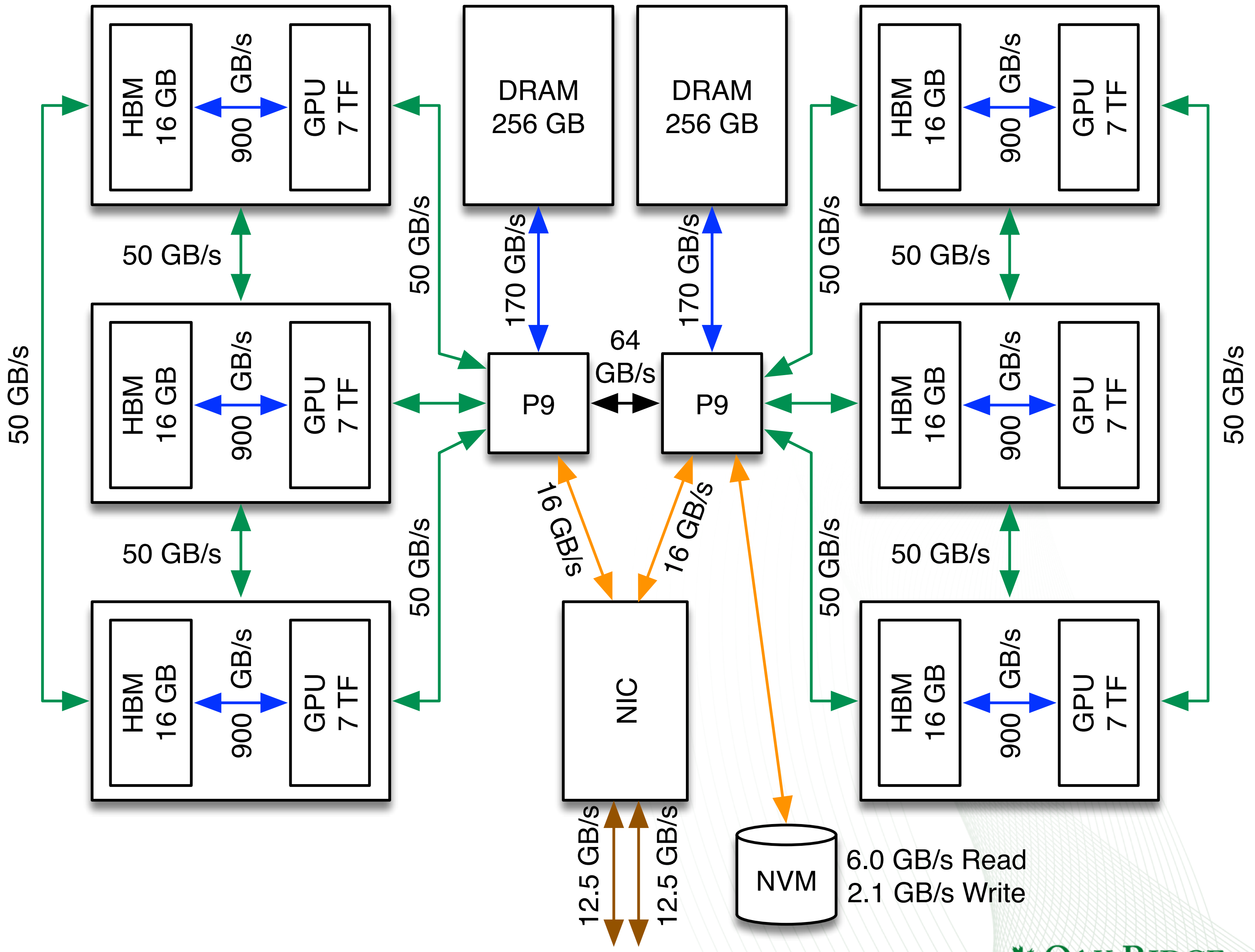
How is Summit Architecture different from Titan? ORNL’s leadership supercomputer

- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and higher memory bandwidth
- Much higher bandwidth between CPUs and GPUs
- Faster interconnect
- Much larger and faster file system
- 7x more performance for only slightly more power (Summit’s 8.8 MW vs Titan’s 8.2)

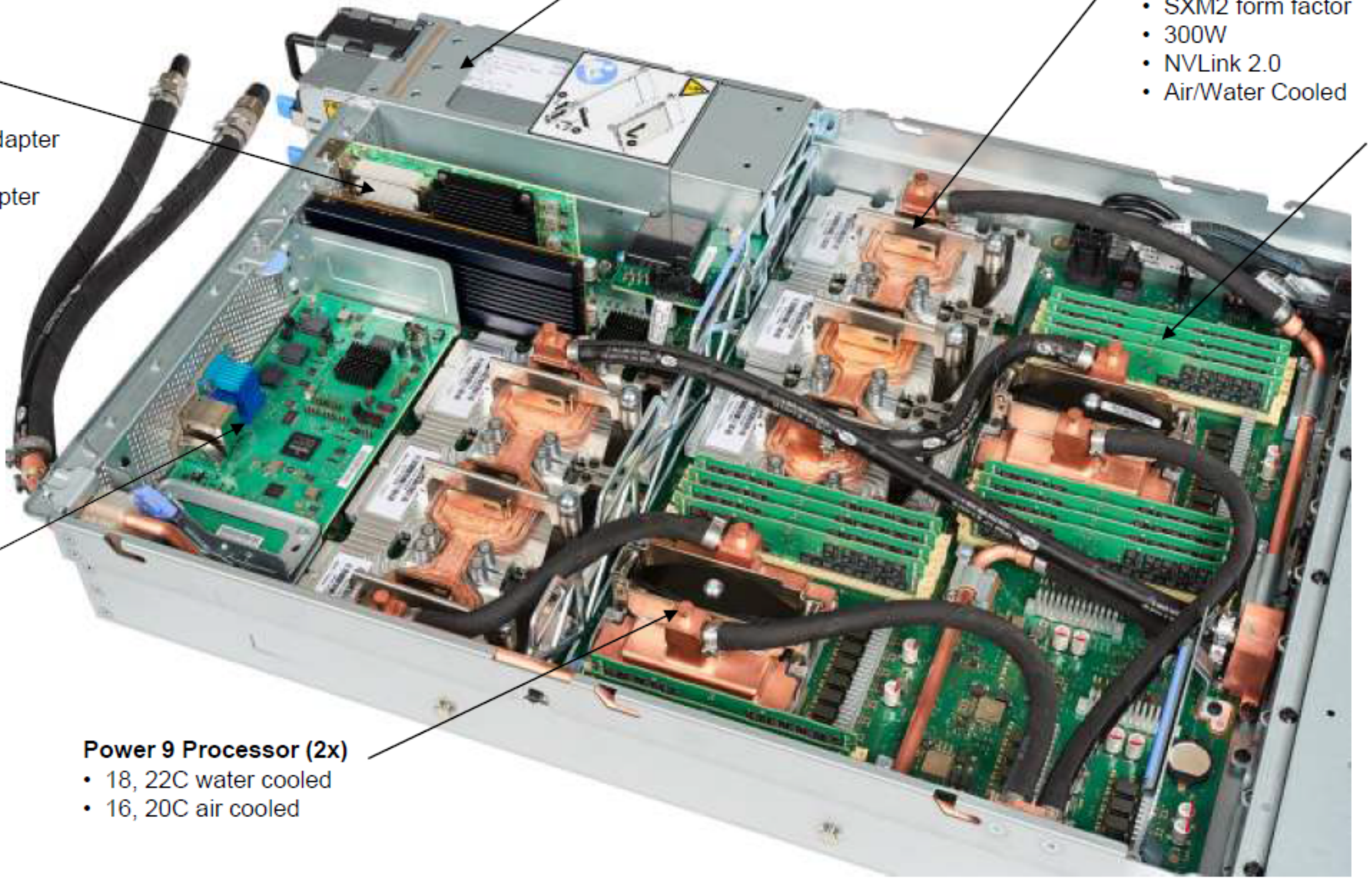
Feature	Titan	Summit
Peak FLOPS	27 PF	200 PF
Max possible Power	9 MW	13 MW
Number of Nodes	18,688	4,608
Node performance	1.4 TF	42 TF
Memory per Node	32 GB DDR3 + 6 GB GDDR5	512 GB DDR4 + 96 GB HBM2
NV memory per Node	0	1.6 TB
Total System Memory	0.7 PB	2.8 PB + 7.4 PB NVM
System Interconnect	Gemini (6.4 GB/s)	Dual Port EDR-IB (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™

Summit Node Schematic

- Coherent memory across entire node
- NVLink v2 fully interconnects three GPUs and one CPU on each side of node
- PCIe Gen 4 connects NVM and NIC
- Single shared NIC with dual EDR ports



Summit Board (1 node) showing the Water Cooling



Power Supplies (2x)
- 2200W
- 200VAC, 277VAC, 400VDC input

NVidia Volta GPU
- 3 per socket
- SXM2 form factor
- 300W
- NVLink 2.0
- Air/Water Cooled

PCIe slot (4x)
- Gen4 PCIe
- 2, x16 HHHL Adapter
- 1, Shared slot
- 1 x8 HHHL Adapter

Memory DIMM's (16x)
- 8 DDR4 IS DIMMs per socket
- 8, 16, 32,64, 128GB DIMMs

BMC Card
- IPMI
- 1 Gb Ethernet
- VGA
- 1 USB 3.0

Power 9 Processor (2x)
- 18, 22C water cooled
- 16, 20C air cooled

Questions?

Summit in Annex Bldg
Titan here
OAK RIDGE
National Laboratory