OAK RIDGE National Laboratory

Summit and **Frontier** at the **Oak Ridge Leadership Computing Facility**

Swaroop Pophale
Programming Models, CSMD
Oak Ridge National Laboratory

August 2, 2021
Argonne Training Program on Extreme-Scale Computing 2021

U.S. DEPARTMENT OF ENERGY

Outline

- OLCF Mission
- OLCF Roadmap to Exascale
- Summit System Overview
- Frontier System Overview

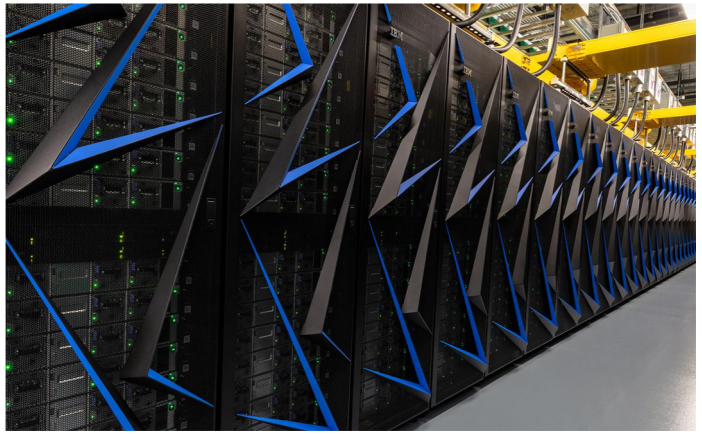
OAK RIDGE National Laboratory

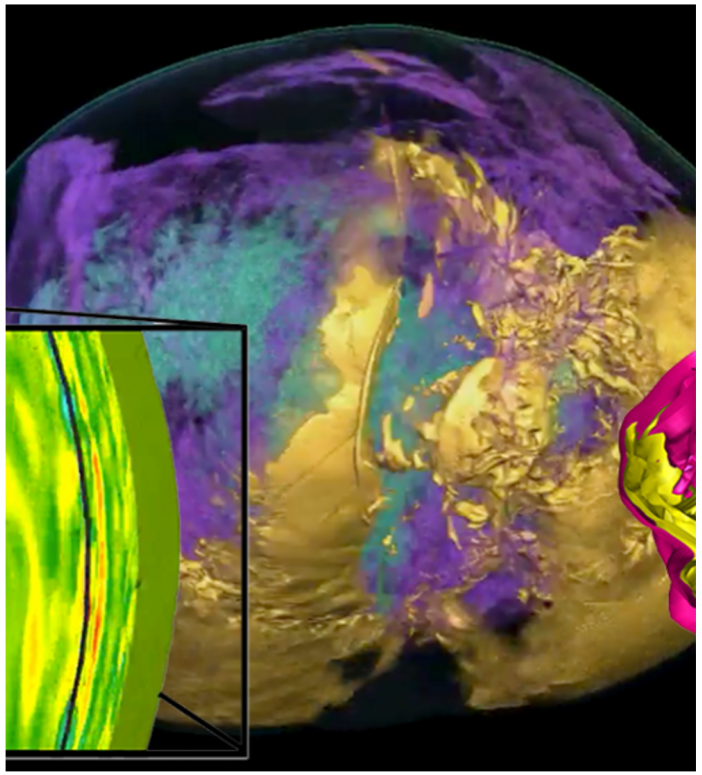
Oak Ridge Leadership Computing Facility (OLCF)

Oak Ridge Leadership Computing Facility (OLCF) Mission

The OLCF is a DOE Office of Science National User Facility whose mission is to enable breakthrough science by:

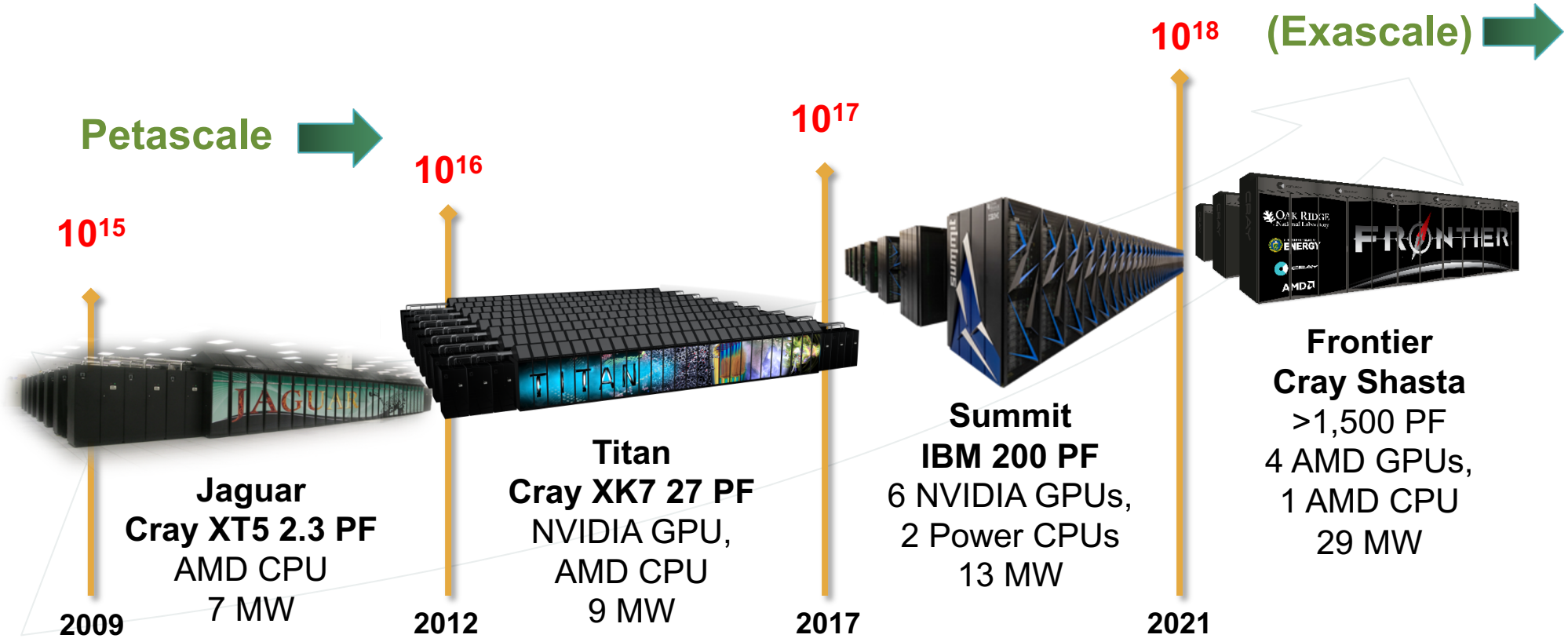
- Fielding the most powerful capability computers for scientific research,
- Building the required infrastructure to facilitate user access to these computers,
- Selecting a few time-sensitive problems of national importance that can take advantage of these systems,
- Partnering with these teams to deliver breakthrough science (Liaisons)





Oak Ridge Leadership Computing Facility Roadmap to Exascale

Mission: Providing world-class computational resources and specialized services for the most computationally intensive global challenges for researchers around the world.



ORNL Summit System Overview

System Performance

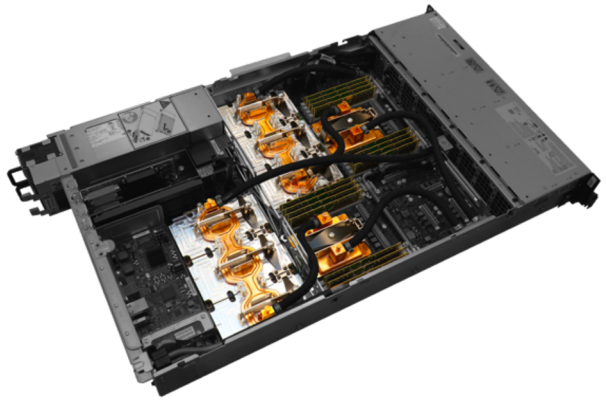
- Peak of 200 Petaflops (FP_{64}) for modeling & simulation
- Peak of 3.3 ExaOps (FP_{16}) for data analytics and artificial intelligence

The system includes

- 4,608 nodes
- Dual-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



Summit Demonstrated Its Balanced Design (2018)

#1 on Top 500, #1 HPCG, #1 Green500, and #1 I/O 500

144 PF HPL
#1 raw performance

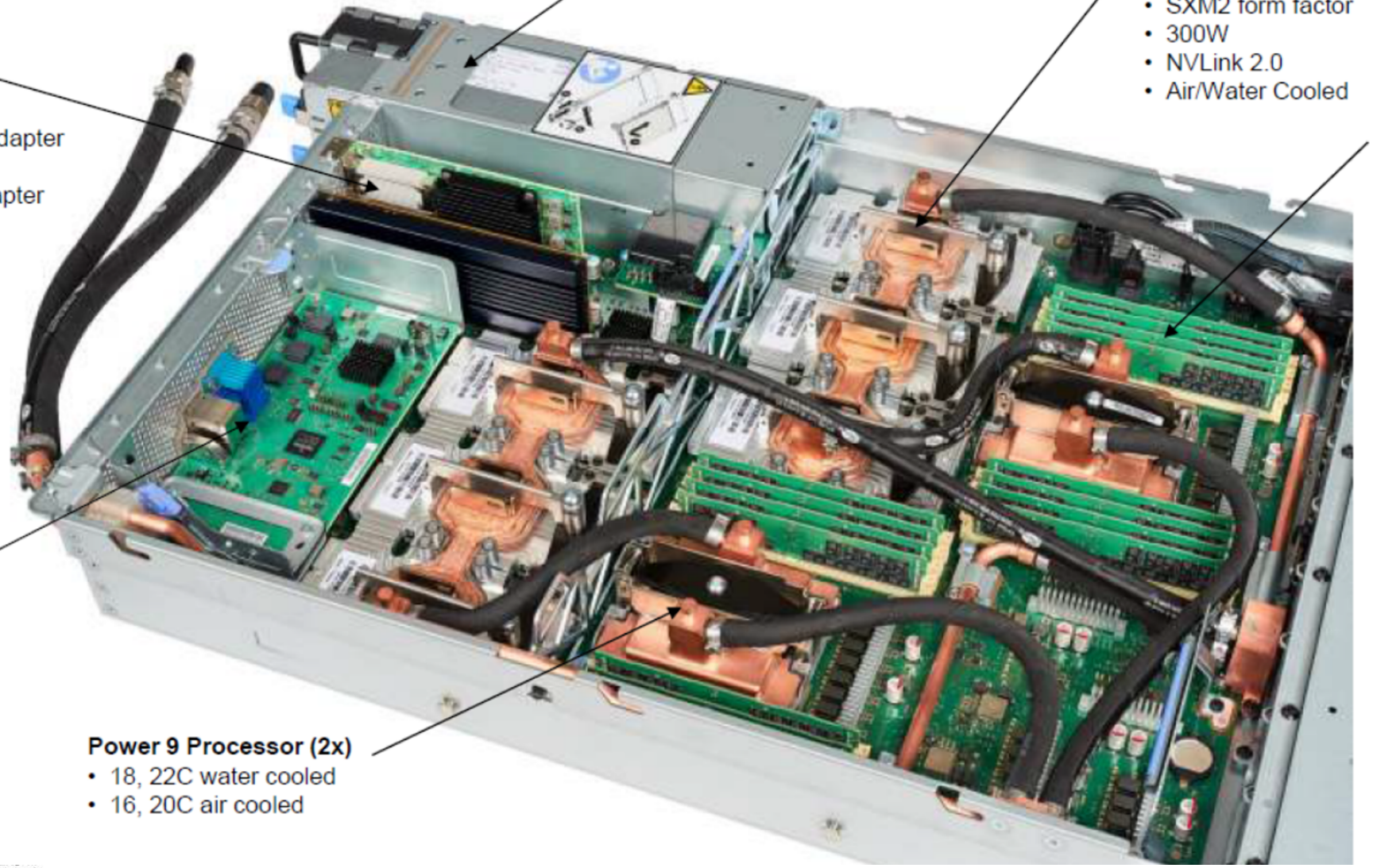
2.9 PF HPCG
#1 fast data movement

14.668 GF/W
#1 energy efficiency

#1 HPC storage performance

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Summit Board (1 node)



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

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

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

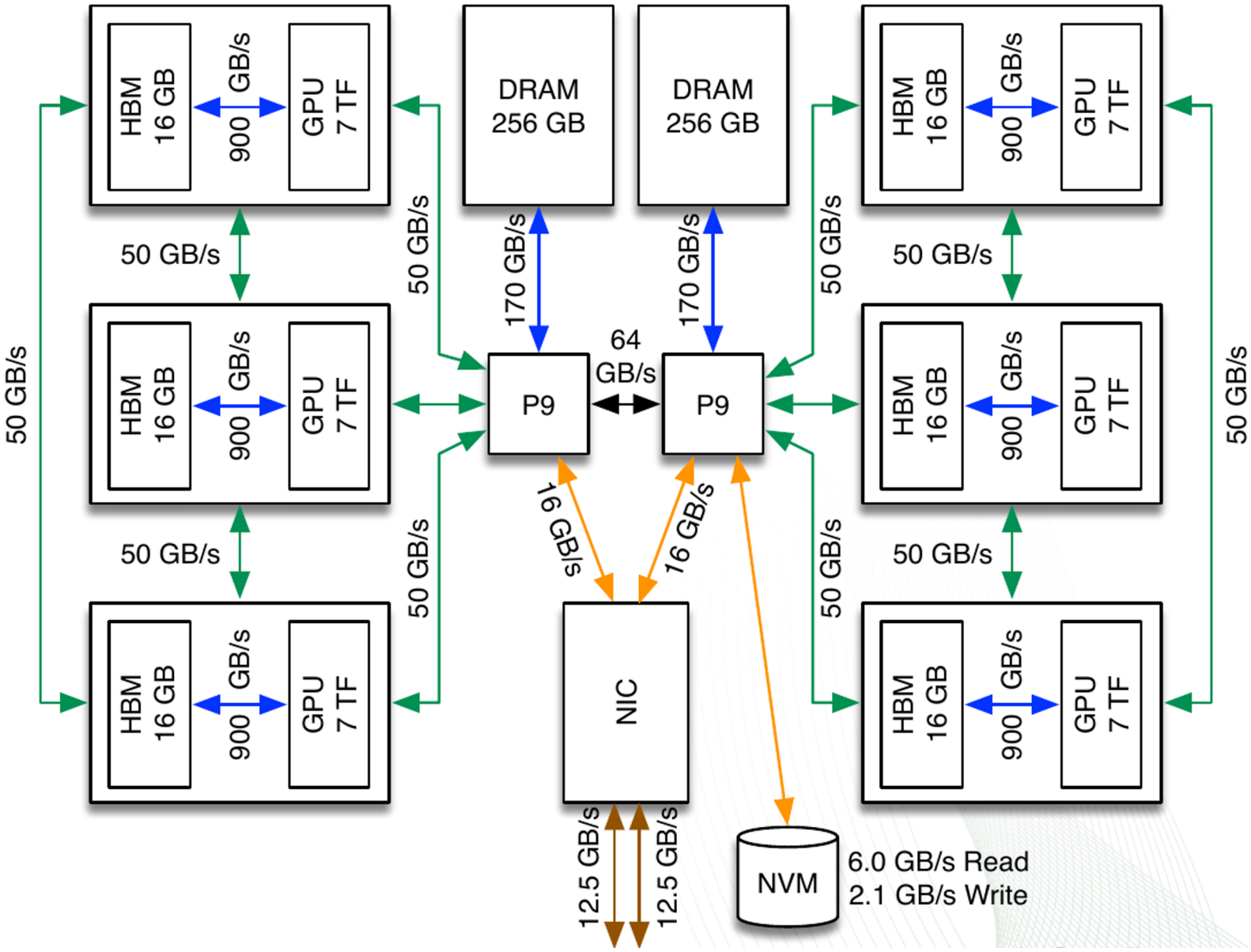
PCIe slot (4x)
- Gen4 PCIe
- 2, x16 HHHL Adapter
- 1, Shared slot
- 1 x8 HHHL Adapter

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

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

Summit Node Schematic

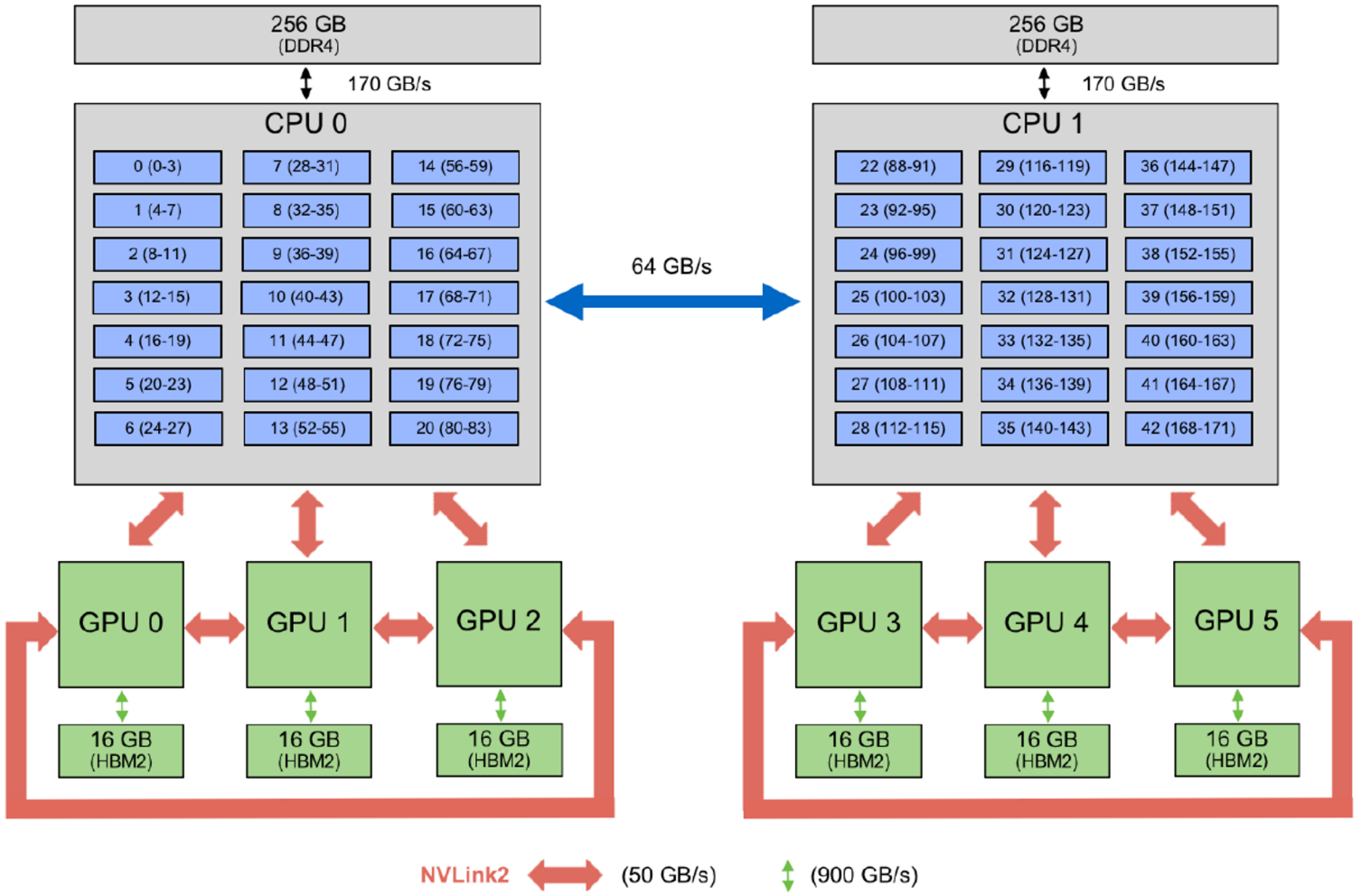
- Coherent memory across entire node
- NVLink v2 fully interconnects three GPUs and one CPU on each side node
- PCIe Gen4 connects NVMe and NIC
- Single shared NIC with dual EDR ports



Summit POWER9 Processors

IBM POWER9 Processor

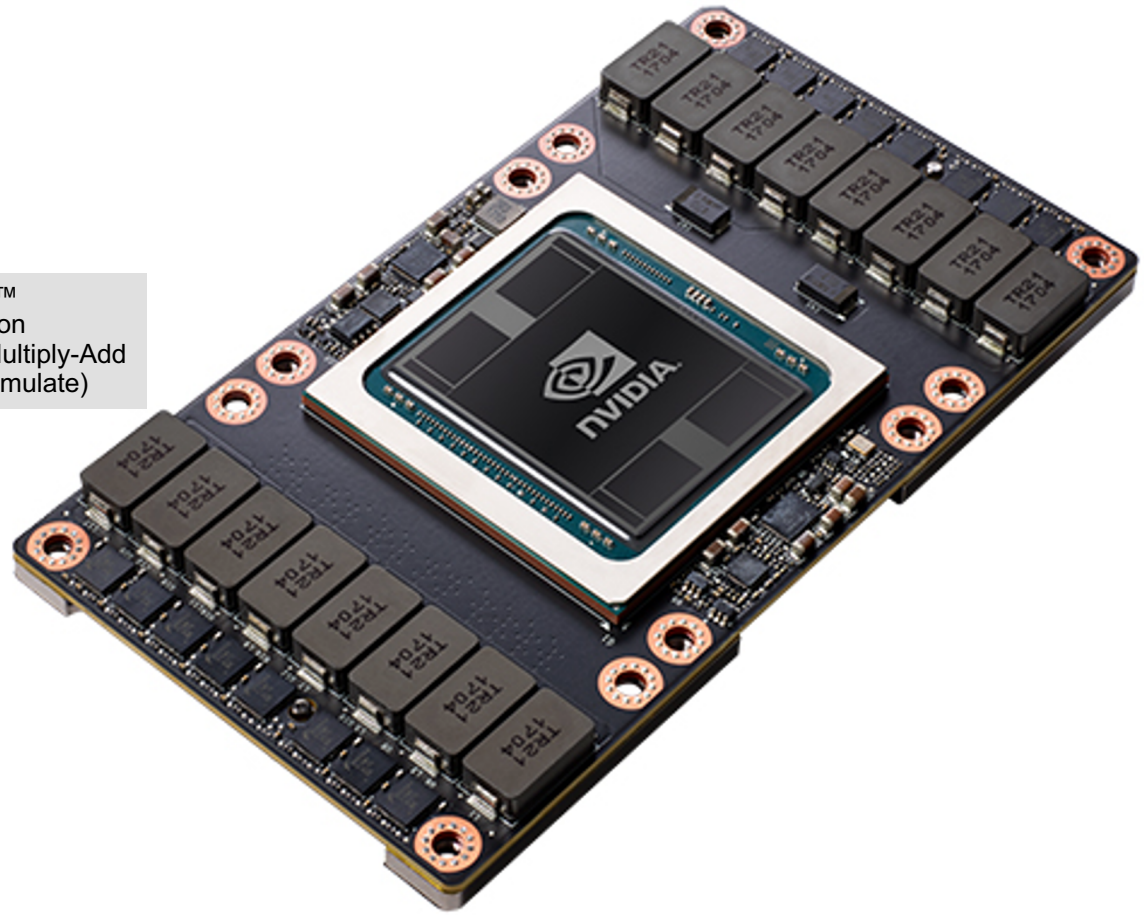
- 22 cores active, 1 core reserved for OS → reduce jitter
- 4 hardware threads (HT) per core
- Three SMT modes: SMT1, SMT2, SMT4. Each thread operates independently.
- 4 HT shares L1 cache, 8 HT (2 cores) shares L2 and L3 cache



Summit GPUs: 27,648 NVIDIA Volta V100s

	Tesla V100 for NVLink	Tesla V100 for PCIe
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 TeraFLOPS SINGLE-PRECISION 15.7 TeraFLOPS DEEP LEARNING 125 TeraFLOPS	DOUBLE-PRECISION 7 TeraFLOPS SINGLE-PRECISION 14 TeraFLOPS DEEP LEARNING 112 TeraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 16 GB HBM2 BANDWIDTH 900 GB/s	

TensorCores™
Mixed Precision
(16b Matrix-Multiply-Add and 32b Accumulate)

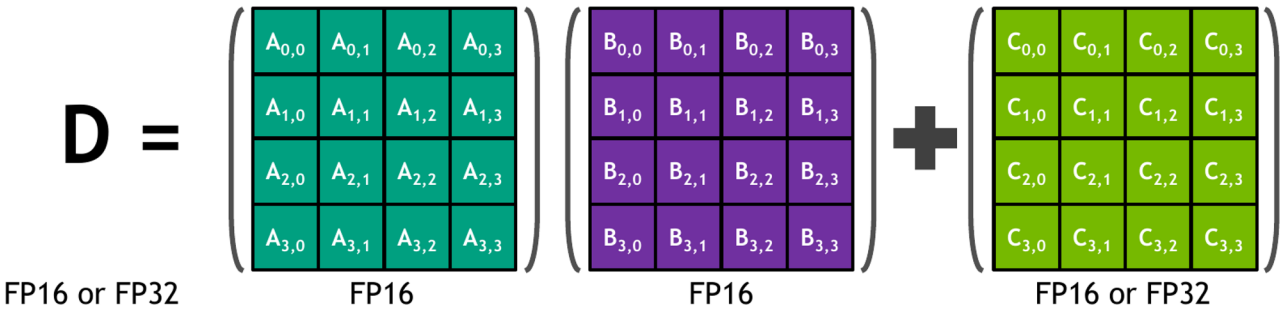


Note: The performance numbers are peak and not representative of Summit's Volta

Summit GPUs: 27,648 NVIDIA Volta V100s (2)

Tensor cores on V100:

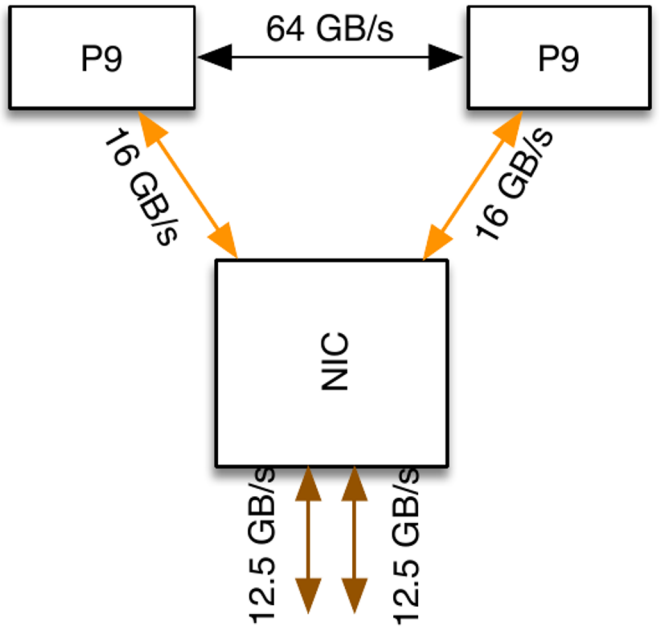
- Tensor cores do mixed precision multiply add of 4x4 matrices



- 640 Tensor cores (8 on each 80 SMs)
- Up to 125 Half Precision (FP_{16}) TFlops
- Requires application to figure out if/when utilizing mixed/reduce precision is possible
  - e.g. see Haidar et al (ICL at UTK), SC18 paper
  - CoMet Comparative Genomics application (2018 ACM Gordon Bell Prize winner), achieving 2.36 ExaOps (mixed-precision) on Summit

Summit Network

- Mellanox EDR Network with non-blocking fat-tree topology
  - Bisection bandwidth 115 TB/s
  - 2 physical ports per node (4 virtual) at 25 GB/s
    - must use both sockets to get full bandwidth
  - Set to minimize latency by default (tune-able)



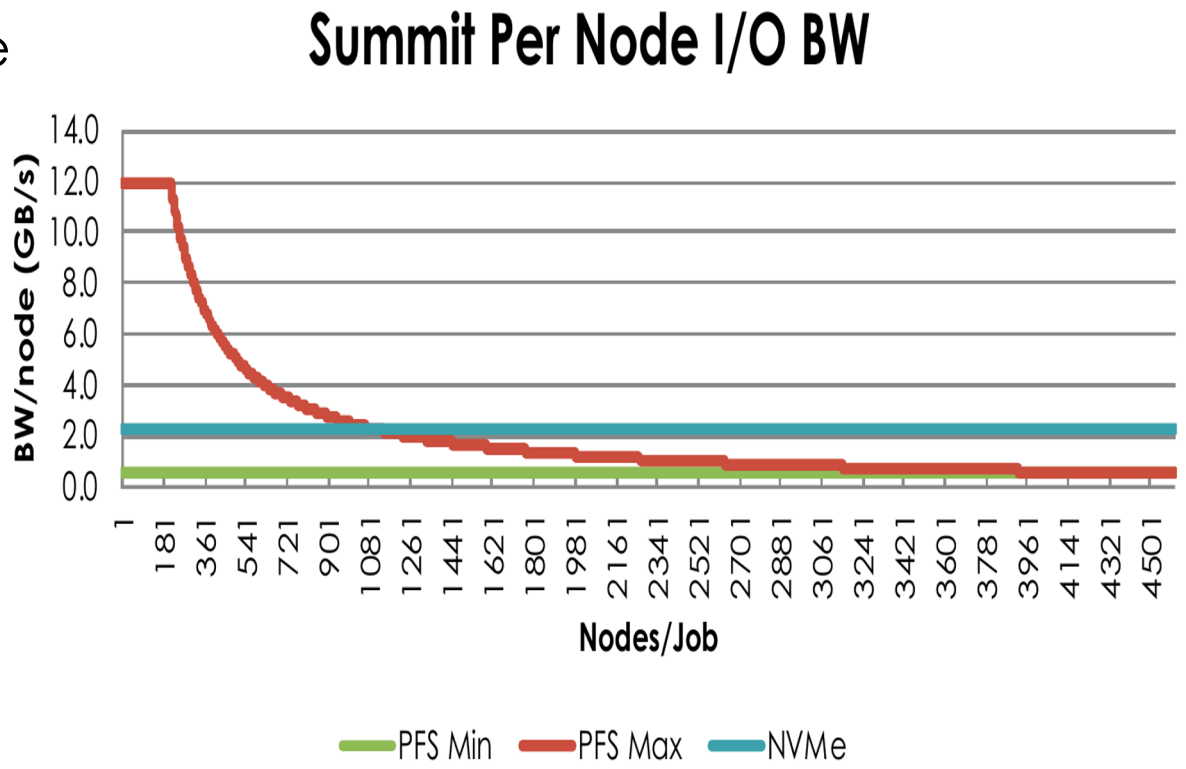
- Adaptive routing
  - Enable bypassing congestions
  - Out of order packets on the network
  - Packets are load balanced at each switch

- Scalable Hierarchical Aggregation (and) Reduction Protocol
  - SHARP: network builds trees in switches to accelerate some collective operations
  - Supported collectives (small <=2048): barrier, broadcast, reduce, allreduce

Summit Parallel File System and Burst Buffers (NVME)

- Alpine "SpectrumScale" File system:
  - 12-14 GB/s per node, 2.5 TB/s aggregate
  - Full system job: ~550 MB/s per node
  - Every node has access to the same space → can support multiple modes: single-shared file, file per rank, etc.

- Node Local NVME:
  - Samsung PMI1725A: Write 2.1 GB/s, Read 5.5 GBs
  - Scales linearly with job size
  - Shared only by ranks on a node,
  - Must drain to PFS at the end of a job (using tools or 'manually')



OAK RIDGE National Laboratory

Summit Programming Environment

Summit Compilers and Programming Model

All compilers (except Clang) support C, C++ and Fortran

Compiler	CUDA (C)	CUDA Fortran	OpenMP 4.5 (offload)	OpenMP (CPU)	OpenACC
PGI	✓	✓		✓	✓
GCC	✓		✓ (*)	✓	✓
IBM XL	✓	✓	✓	✓	
LLVM (C & C++)	✓		✓	✓	

*: functional only

Summit Debuggers and Performance Tools

Debugger
DDT
Valgrind
GDB

Performance Tools
Open|SpeedShop
TAU
HPCToolkit (IBM)
HPCToolkit (Rice)
VAMPIR
NVIDIA Nsight
Score-P

Summit Numerical Library

Library	OSS or Proprietary	CPU Node	CPU Parallel	GPU
IBM ESSL	Proprietary	✓		✓
FFTW	OSS	✓	✓	✓
ScaLAPACK	OSS	✓	✓	
PETSc	OSS	✓	✓	
Trilinos	OSS	✓	✓	✓*
BLAS-1, -2, -3	Proprietary (thru ESSL)	✓		✓
NVBLAS	Proprietary			✓
cuBLAS	Proprietary			✓
cuFFT	Proprietary			✓
cuSPARSE	Proprietary			✓
cuRAND	Proprietary			✓
Thrust	Proprietary			✓

Summit Job Launcher: `jsrun`

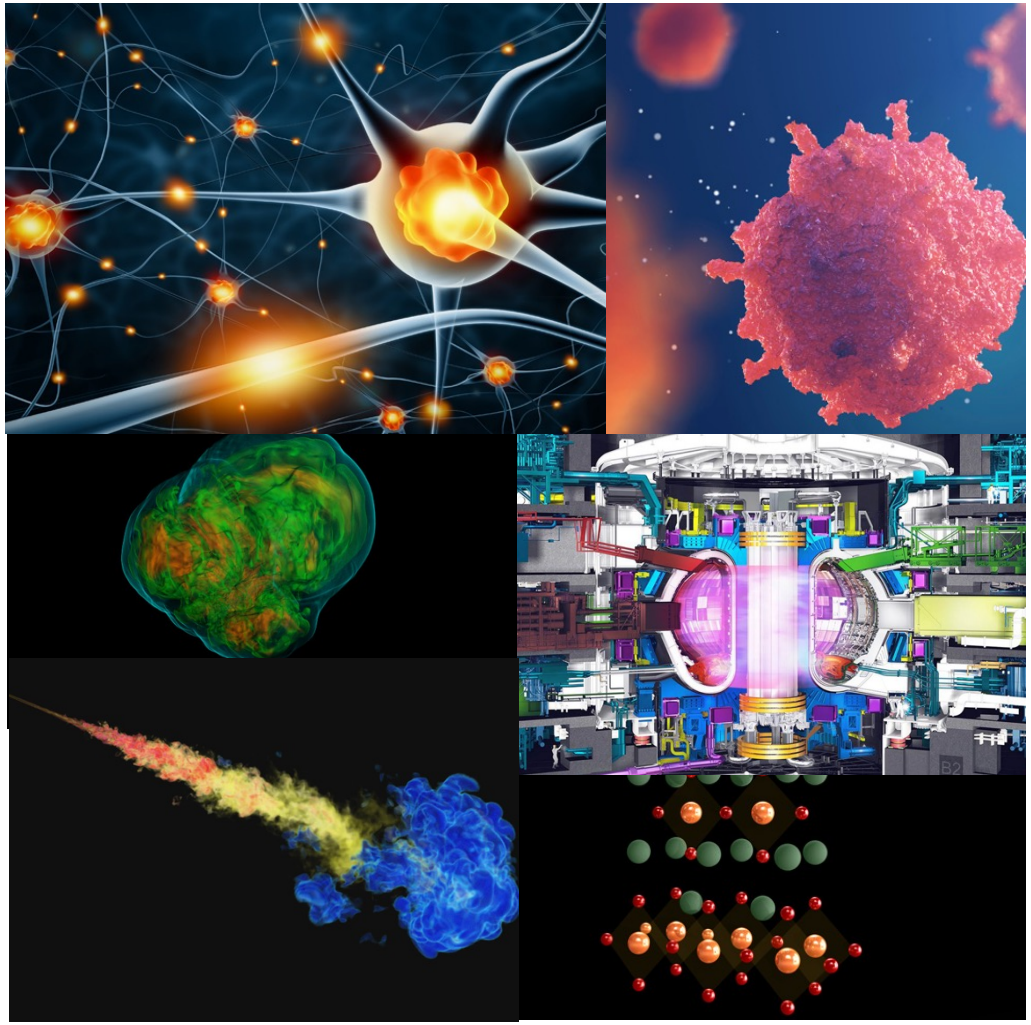
- jsrun provides abstraction of a node with a concept of 'resource set'
  - motivated by the fact that Summit has powerful "fat" nodes
- Resource set:
  - sub group of resources (GPUs, CPUs) within a node
  - Implemented using cgroup under the hood
  - executes <N> MPI processes (with threads) and manages placement
- Node-sharing (e.g. multiple executables) is possible within a job:
  - Multiple Programs Multiple Data (MPMD)
  - concurrently execute compute intensive GPU-only job with CPU-only data analysis / visualization

Programming Multiple GPUs

- Multiple paths, with different levels of flexibility and sophistication, e.g.:
  - Simple model: 1 MPI or 1 thread per GPU
  - Sharing GPU: multiple MPIs or threads share a GPU
  - Single MPI using multiple GPUs
  - Expose the node-level parallelism directly: multiple processes multiple GPUs
- Exposing more (node-level) parallelism is key to scalable applications from petascale-up

Summit and Scientific Discovery

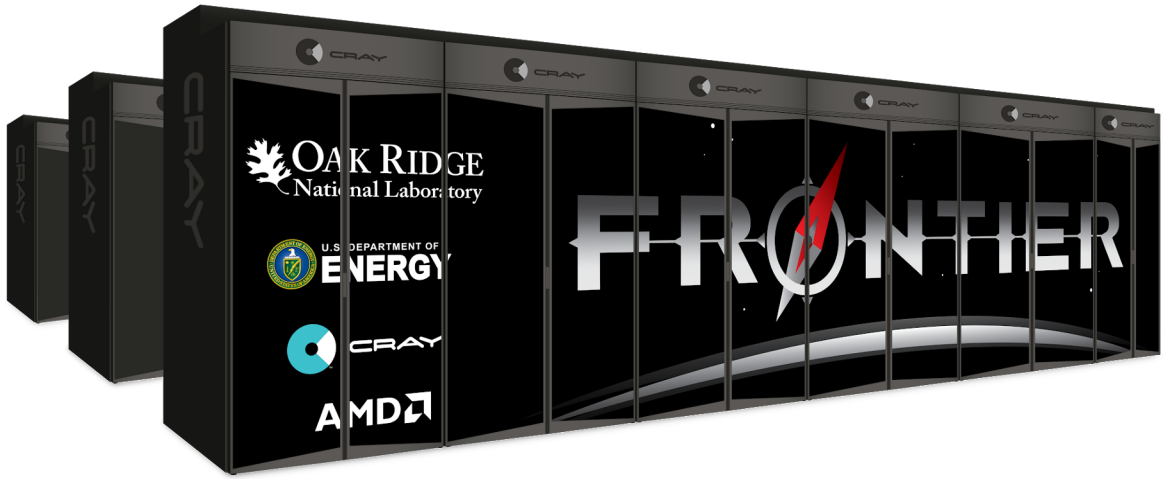
- Deep Learning for
  - Human System Biology
  - Cancer Research
- Plasma Fusion (XGC)
- Combustion (RAPTOR)
- Astrophysics (Flash)
- Materials (QMCPACK)




FRONTIER
DIRECTION OF DISCOVERY
ORNL's exascale supercomputer designed to deliver world-leading performance in 2021.

Frontier Overview

- Partnership between ORNL, Cray, and AMD
- Frontier will be delivered in late 2021
- Peak performance greater than 1.5 EF
- More than 100 Cray Shasta cabinets



Frontier Node Architecture

- An AMD EPYC(™) processor with four Radeon Instinct(™) GPU accelerators purpose-built for exascale computing
- Fully connected with high-speed AMD Infinity Fabric links
- Coherent memory across the node
- 100 GB/s injection bandwidth
- Near-node NVM storage

Frontier I/O subsystem

- Will consist of two major components:
  - an in-system storage layer and
  - a center-wide file system called Orion
- Orion will use open-source Lustre and ZFS technologies
  - Lustre for distributed name space, data on metadata, and progressive file layouts
  - 40 Lustre metadata server nodes and 450 Lustre object storage service (OSS) nodes.
  - Each OSS node will provide one object storage target (OST) device for performance and two OST devices for capacity

Frontier I/O subsystem (1)

- Orion will comprise three tiers:
  - 5,400 nonvolatile memory express (NVMe) devices; 11.5 petabytes (PB) of capacity at peak read-write speeds of 10 TBps; more than 2 million random-read IOPS;
  - hard-disk-based capacity 679 PB of capacity at peak read speeds of 5.5 TBps and peak write speeds of 4.6 TBps with more than 2 million random-read IOPS; and
  - metadata tier of 480 NVMe devices providing an additional capacity of 10 PB.

Frontier Programming Environment

Vendor-Provided

- Cray Programming Environment (CPE)
  - Includes Cray compiler for C, C++, and Fortran plus GCC compiler. All the Cray profiling, tuning, and debugging tools. OpenMP and Cray MPI optimized for AMD GPU direct.
- AMD ROCm programming environment
  - Includes LLVM compiler to generate optimized code for both the AMD Epyc CPU and Instinct GPU. It will support: C, C++, and Fortran and have GPU offload support. HIP for converting CUDA codes to run on AMD GPUs.

Other Sources

- ECP
  - LLVM enhancements: Flang (Fortran front-end), OpenMP, OpenACC
  - Kokkos and RAJA
  - HIP LZ (HIP support for Aurora)
  - MPI, HPCToolkit, PAPI enhancements
  - …
- ALCF + OLCF
  - Pilot implementation of DPC++/SYCL for Frontier
- OLCF
  - GCC enhancements to better support OpenACC, OpenMP, Fortran on Summit and Frontier

Frontier Programming Environment (1)

Items in green are also available on Summit

- Compilers Offered
  - Cray PE (C/C++ LLVM-based; Cray Fortran)
  - AMD ROCm (LLVM-based)
  - GCC
- Programming Languages & Models Supported (in which compilers)
  - C, C++, Fortran (all)
  - OpenACC (GCC) 2.6 substantially complete, 2.7 planned
  - OpenMP (all)
  - HIP (Cray, AMD) – New: Cray has added HIP support to CPE
  - Kokkos/RAJA (all)
  - UPC (Cray, GCC)
- Transition Paths
  - CUDA: semi-automatic translation to HIP
  - CUDA Fortran: HIP kernels called from Fortran (a more portable approach)
    - CUDA Fortran kernels need to be translated to C++/HIP (manual process)
    - Fortran bindings to HIP and ROCm libraries and HIP runtime available through AMD's hipfort project

Frontier Programming Environment Migration Path

- HIP (heterogenous-compute Interface for Portability) is an API developed by AMD for portable code on AMD and NVIDIA GPU
  - uses CUDA or ROCm under the hood
- The API is very similar to CUDA
- AMD has developed a “hipify” tool to convert from CUDA to HIP
- HIP is available on Summit and is updated regularly

Frontier Programming Tools

Debuggers and Correctness Tools

Tool
System-Level Tools
Arm DDT
Cray CCDB
Cray ATP
STAT
Node-Level Tools
ROCgdb
Cray GDB4HPC

Performance Tools

Tool
System-Level Tools
Arm MAP/Performance Reports
CrayPat/Apprentice2 (Cray)
Reveal (Cray)
TAU
HPCToolkit
Score-P / VAMPIR
Node-Level Tools
gprof
PAPI
ROCprof
ROC-profiler & ROC-tracer libraries

Items in green are also available on Summit

Frontier Scientific Libraries and Tools

Functionality	CPU	GPU	Notes
BLAS	**Cray LibSci**, **AMD BLIS**, PLASMA	**Cray LibSci_ACC**, **AMD roc/hipBLAS**, AMD rocAMD ROCm Tensile, **MAGMA**	MAGMA and PLASMA are open source software led by the UTK Innovative Computing Laboratory
LAPACK	**Cray LibSci**, **AMD libFlame**, PLASMA	**Cray LibSci_ACC**, **AMD roc/hipSolver**, **MAGMA**	
ScaLAPACK	Cray LibSci	ECP SLATE, Cray LibSci_ACC	
Sparse		**AMD roc/hipSparse**, AMD rocALUTION	
Mixed-precision iterative refinement	Cray IRT, MAGMA	MAGMA	
FFTW or similar	**Cray**, **AMD**, ECP FFTX, FFT-ECP	**AMD rocFFT**, ECP FFTX, FFT-ECP	FFT-ECP focuses on 3D FFTs
PETSc, Trilinos, HYPRE, SUNDIALS, SuperLU			Spack recipes from ECP xSDK

Functionality in **green** is also available on Summit

Frontier Timeline

- Early Access System (spock) now available
  - “n-1” hardware (processors, network, etc.)
  - With the evolving Cray and AMD programming environments
- Frontier will be delivered in 2021, with acceptance expected in first half of 2022
  - ECP expected to gain access in June 2022
  - INCITE access will ramp up from Jan 2023 to full allocation starting Jan 2024
  - ALCC access will ramp up from Jul 2023 to full allocation starting Jul 2024

In the mean time

- Summit provides many of the same tools and a similar architecture
  - Especially useful if you’re new to GPU programming
- Early Access systems will provide the (evolving) software stack on near-Frontier hardware

System Comparisons: Titan, Summit, and Frontier

System	Titan (2012)	Summit (2017)	Frontier (2021)
Peak	27 PF	200 PF	> 1.5 EF
# nodes	18,688	4,608	> 9,000
Node	1 AMD Opteron CPU 1 NVIDIA Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 AMD EPYC CPU 4 AMD Radeon Instinct GPUs
Memory		2.4 PB DDR4 + 0.4 HBM + 7.4 PB On-node storage	4+ PB DDR4 + 4+ PB HBM2e + 35+ PB On-node storage, 75 TB/s Read 38 Write
On-node interconnect	PCI Gen2 No coherence across the node	NVIDIA NVLINK Coherent memory across the node	AMD Infinity Fabric Coherent memory across the node
System Interconnect	Cray Gemini network 6.4 GB/s	Mellanox Dual-port EDR IB 25 GB/s	Four-port Slingshot network 100 GB/s
Topology	3D Torus	Non-blocking Fat Tree	Dragonfly
Storage	32 PB, 1 TB/s, Lustre Filesystem	250 PB, 2.5 TB/s, IBM Spectrum Scale™ with GPFS™	Lustre with: 679 PB HDD+11 PB Flash Performance Tier at 10 TB/s (R/W) and 10 PB Metadata Flash
Power	9 MW	13 MW	29 MW

Acknowledgments

- The OLCF team

This work was performed under the auspices of the U.S. DOE by Oak Ridge Leadership Computing Facility at ORNL under contracts DEAC05-00OR22725

Resources

- More info on Summit:
  - Summit user guide: https://www.olcf.ornl.gov/for-users/system-user-guides/summit/
  - OLCF training archive: https://www.olcf.ornl.gov/for-users/training/training-archive/
  - Vazhkudai, *et. al.* The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems. SC18 Proceedings.

- For latest on Frontier:
  - https://www.olcf.ornl.gov/frontier/

Thanks

E-mail: pophaless@ornl.gov