Hewlett Packard Enterprise

Frontier System Architecture

Joe Glenski, Sr. Distinguished Technologist

February 15, 2023

Topics

- Frontier System Overview
- Cabinet and Blade Design
- Node Design
- HPE Slingshot Interconnect
- Login Nodes
- Storage
- Application Software Stack

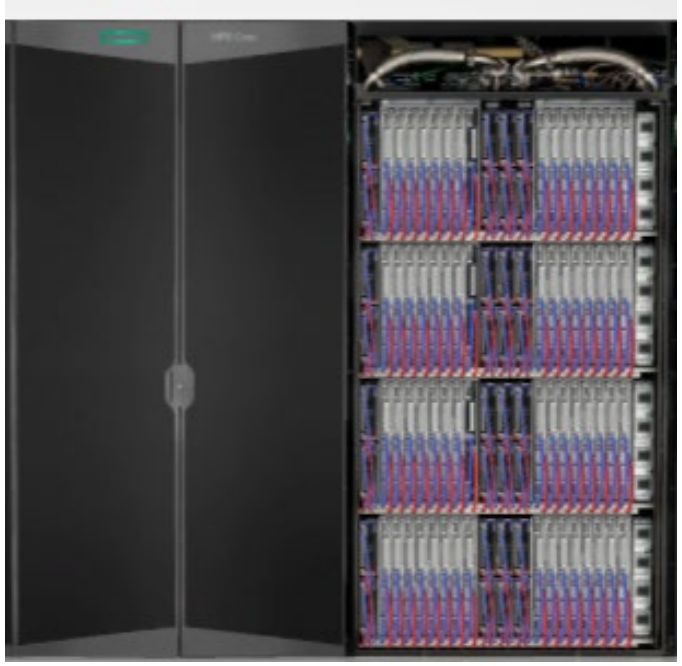
Frontier System Overview

HPE Cray EX Supercomputer architecture

- Dense liquid cooled compute cabinets
- 74 cabinets holding 9408 compute nodes
- Double-precision performance of 1.1 exaflops
- Standard racks holding support and management nodes
- AMD 64-core Optimized 3rd Gen EPYC CPUs
- AMD MI250X Instinct GPUs
- HPE Slingshot interconnect with 200 GbE interfaces
- HPE Cray Software stack and AMD ROCm stack
- 679 PB multi-tier Lustre filesystem “Orion”
- Access to OLCF Storage (/ccs/home, etc.)

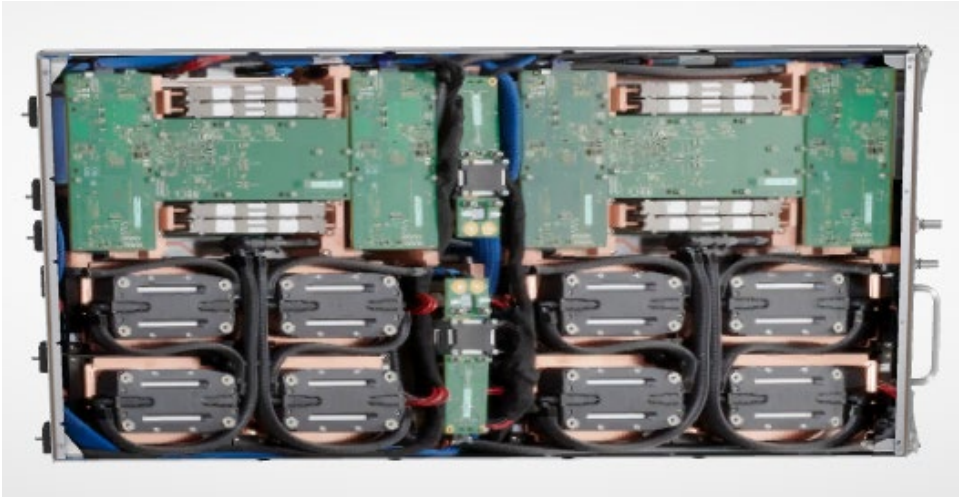
HPE Cray EX4000 Cabinet

- Up to 64 compute blades within 8 compute chassis per rack
- Direct liquid cooling using a shared CDU
- 4 power shelves with a max of 8 rectifiers each
- Provides a very efficient system (FLOPs/Watt)



HPE Cray EX235A Blade

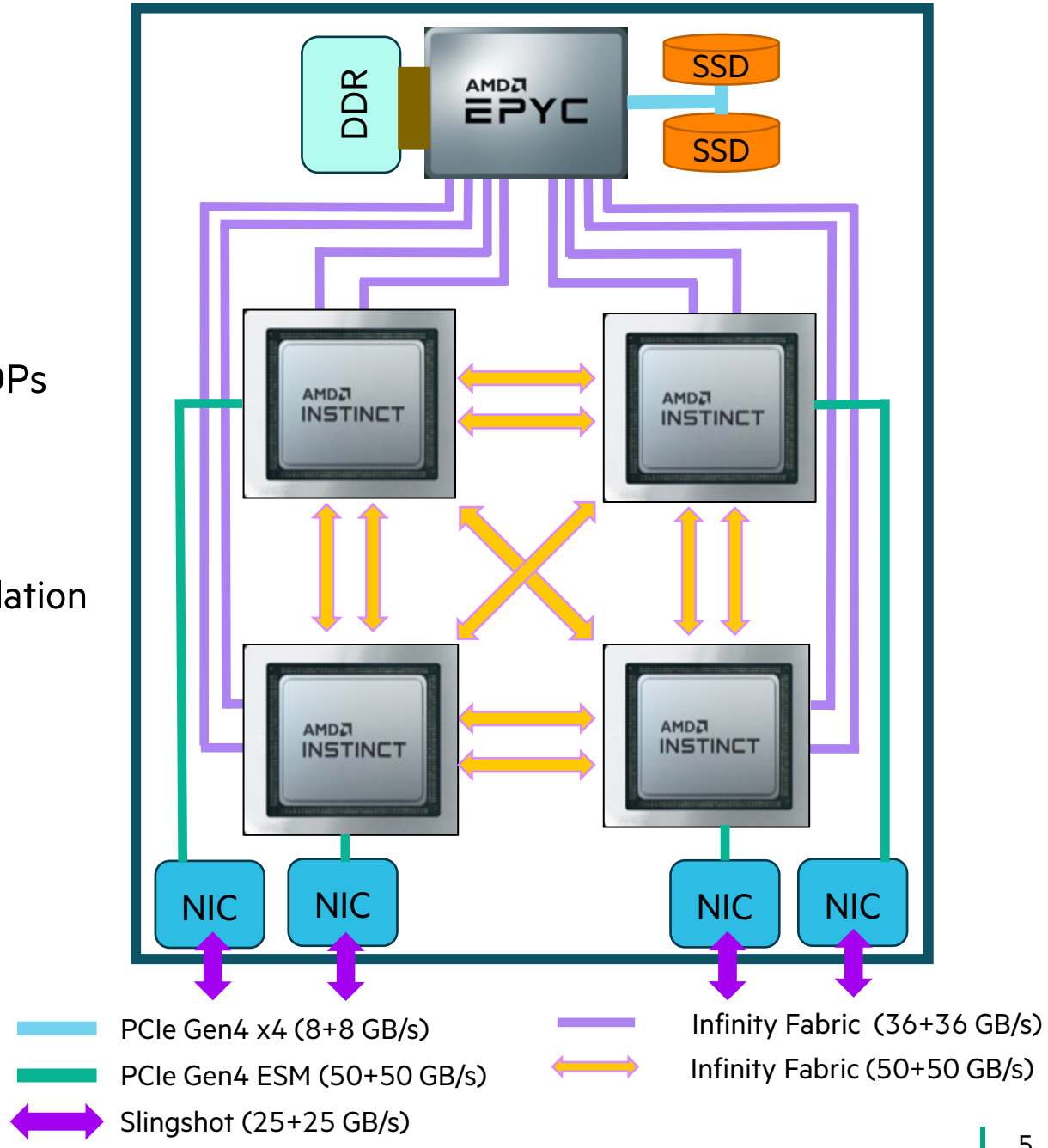
- Two compute nodes per blade
- Each with one CPU and four MI250X accelerators
- Direct liquid cooled for all components
- Supports high power processors > 500W



https://www.hpe.com Then search for “HPE Cray EX Supercomputer”

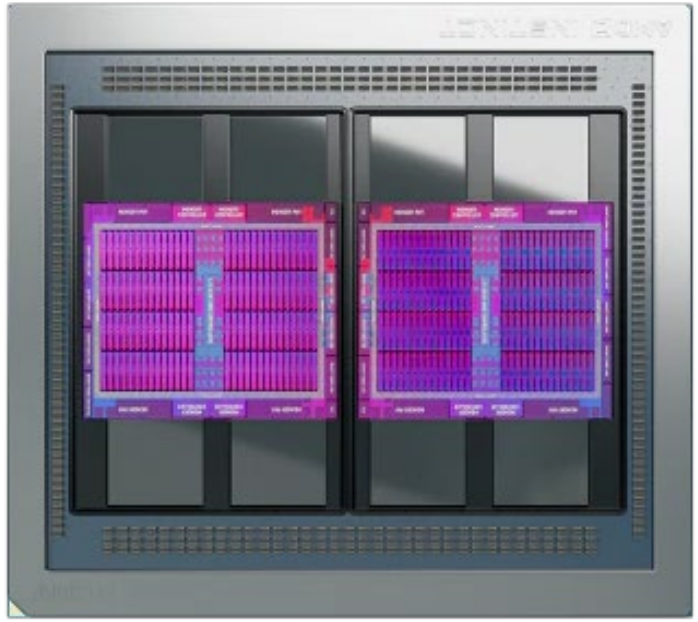
Frontier Compute Node Design

- 1x AMD Optimized 3rd Gen EPYC 64 core processor
  - 2 hardware threads per physical core,
  - 2.0GHz base clock, 3.7GHz boost clock
- 512 GB DDR4 memory with 205 GB/s peak bandwidth
- 2x NVMe 2TB SSDs, peak 8 GB/s R, 4 GB/s W, >1.5M IOPs
- 4x AMD MI250X Instinct GPUs
  - 128 GB High-Bandwidth Memory (HBM2E)
  - 3.2 TB/s peak bandwidth
  - 53 TFLOPS double-precision peak for modeling & simulation
  - 2 Graphic Compute Dies (GCDs)
- AMD Infinity Fabric between CPU and GPUs
  - Peak host-to-device (H2D) and device-to-host (D2H) data transfers of 36+36 GB/s per link
- AMD Infinity Fabric between MI250Xs
  - Peak device-to-device bandwidth of 50+50 GB/s per link, low latency
- Coherent memory between CPU and all GPUs
- 4x HPE Slingshot Interconnect 200 GbE NICs
  - Provides 100 GB/s to other nodes, 25 GB/s per port



MI250X details

- The AMD MI250X has two Graphic Compute Dies (GCDs) per module
- This gives a total of 8 GCDs per node
- The 8 GCDs show as 8 separate GPUs to the OS, Slurm, and ROCm
- Generally easier to refer to the 8 GCDs as GPUs
- And arrange programs to run on nodes with 8 GPUs
- Each Node then has 8 GPUs with the following specifications:
  - HBM Capacity: 64 GB
  - HBM Peak Bandwidth: 1.6 TB/S
  - Compute Units: 110
  - 26.5 TFLOPS double-precision peak
- The 8 GPUs are each associated with one of the 8 CPU L3 cache regions
- All 8 GPUs are connected to each other and to the CPU via AMD Infinity Fabric links
- The two GCDs in the same MI250X have a higher bandwidth Infinity Fabric connection between them, with 200 GB/s peak



AMD MI250X with multiple dies

https://www.amd.com/en/products/server-accelerators/instinct-mi250x

Frontier Compute Node Diagram

https://docs.olcf.ornl.gov/systems/frontier_user_guide.html

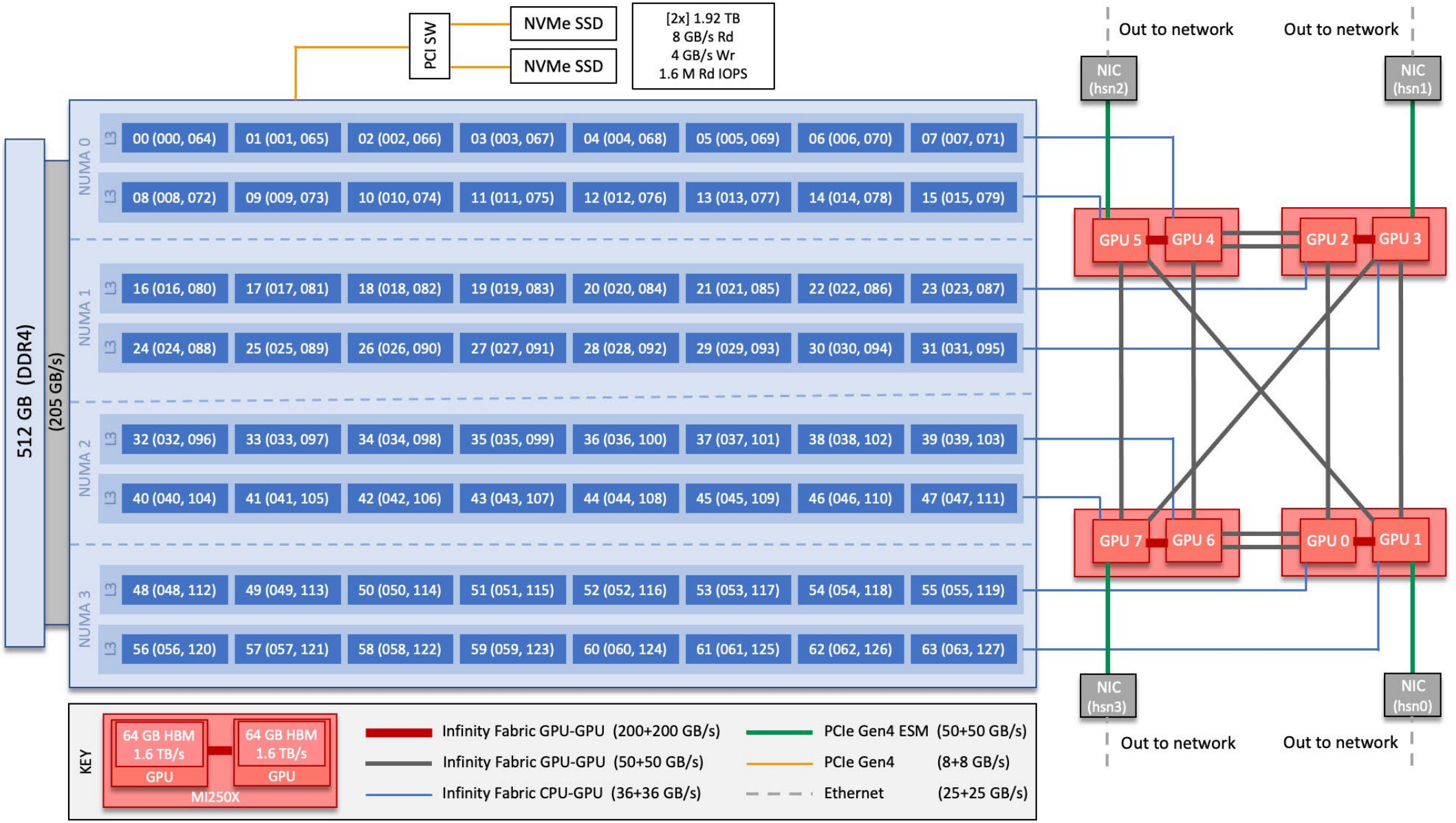


Image from OLCF Frontier User Guide

HPE Slingshot Interconnect

- High speed, low latency network architecture
- Uses proven Dragonfly topology
  - Highly scalable, cost efficient, copper and optical cables
- HPE Slingshot Interconnect switches
  - High radix, 64-port, 12.8 Tb/s bandwidth switch
  - Bi-directional bandwidth of 25 GB/s per port
- HPE Slingshot 200 GbE Interfaces, multiple per node
  - Bi-directional bandwidth of 25 GB/s per link
- Ethernet standards and protocols, plus optimized HPC functionality
- Link level retry and low-latency forward error correction
- Standardized, open API management interfaces
- Advanced flow control features designed to explicitly address congestion and bottlenecks
  - Adaptive Routing, Quality of Service, Congestion Control

D. De Sensi, S. Di Girolamo, K. H. McMahon, D. Roweth and T. Hoefler, An In-Depth Analysis of the Slingshot Interconnect, SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, 2020, pp. 1-14

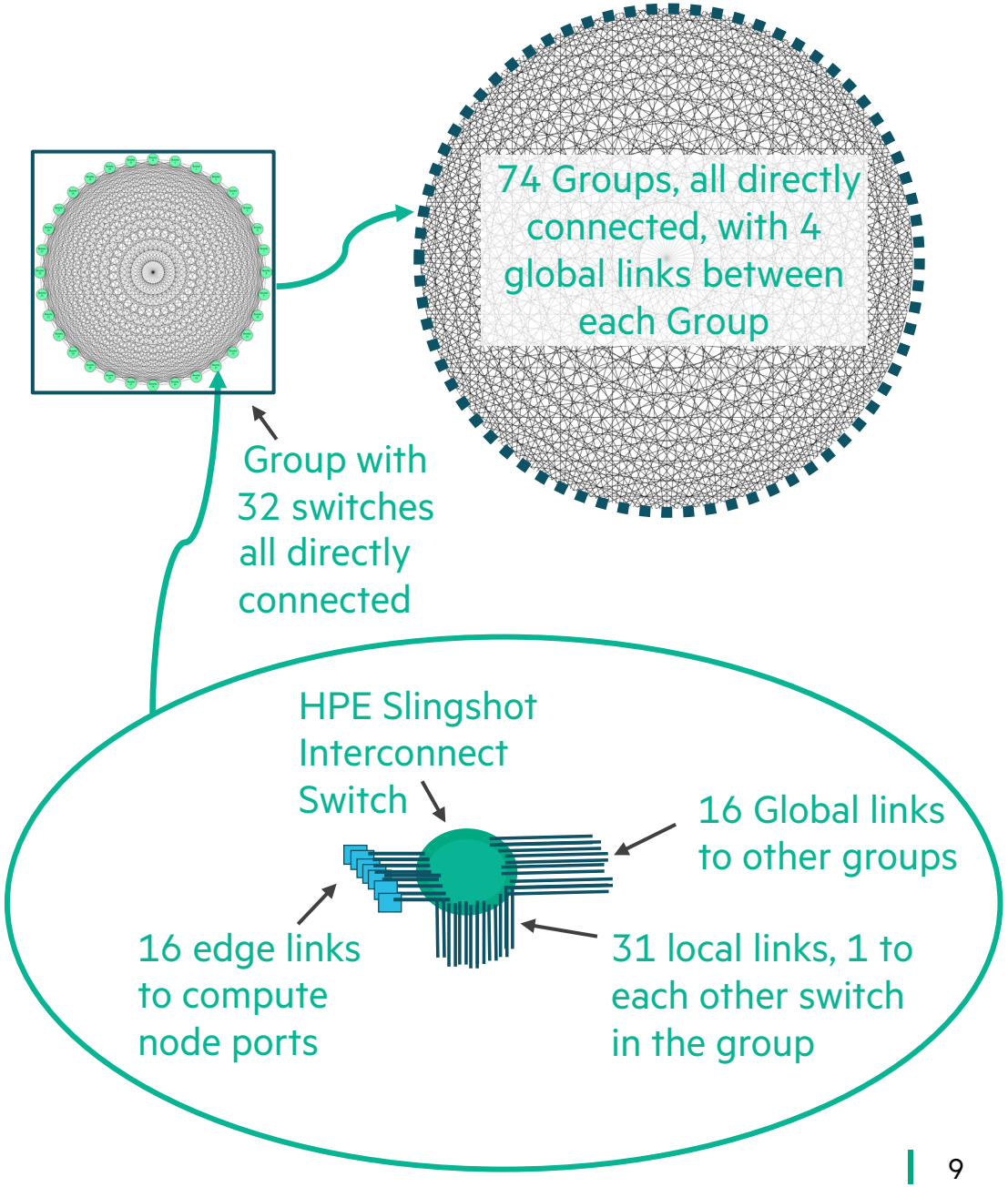
Kim, J., Dally, W., Scott, S., Abts, D.: Cost-Efficient Dragonfly Topology for Large-Scale Systems. IEEE Micro. 29(1), 33–40 (2009)

J. Kim, W. J. Dally, S. Scott, and D. Abts. Technology-driven, highly-scalable dragonfly topology. ACM SIGARCH, 2008.

https://www.hpe.com/us/en/compute/hpc/slingshot-interconnect.html

Frontier HPE Slingshot Topology

- Frontier has 74 compute groups, each with 32 switches
  - High radix, 64-port, 12.8 Tb/s bandwidth switch
- Each group has 128 compute nodes
- All to All connections between groups
  - 4 links between each of the 74 compute groups on Frontier
  - Bi-directional bandwidth of 25 GB/s per link
- Switch connections:
  - Up to 16 Edge links (L0) to compute node ports
  - 31 Local links (L1) to the other switches in the group
  - Up to 16 Global links (L2) to switches in other groups
- Node connections – 4 links from each node to switches
  - HPE Slingshot 200 GbE NICs
  - Bi-directional bandwidth of 25 GB/s per link
  - 4 Ports give total node injection bandwidth of 100 GB/s



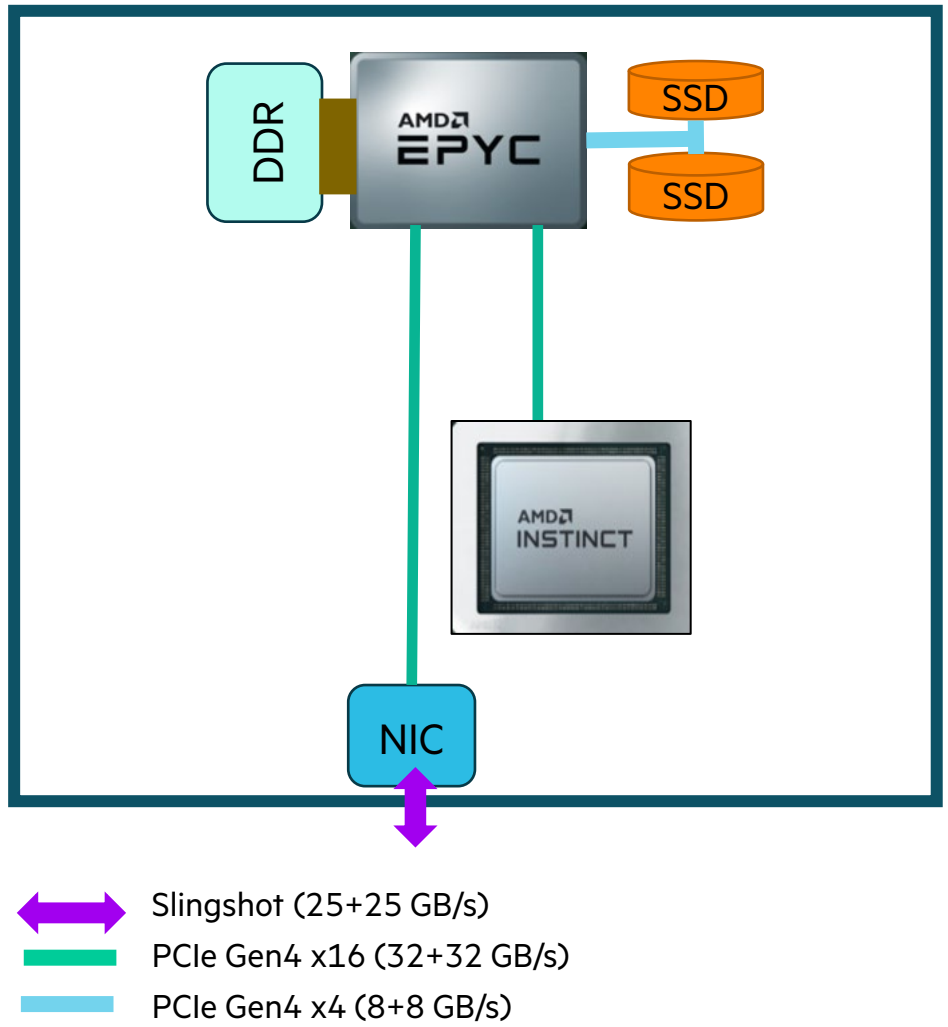
Login nodes, Gateway servers, Lustre servers, etc. are in separate support and storage groups

Frontier Login Nodes

- Frontier has 16 Login Nodes (not all in use)
- These are for users to write/edit/compile code, manage data, submit jobs
- Multiple users share these nodes

- Each login node has:
  - 1x AMD EPYC 7763 (Milan) CPU
    - 64 Cores, 128 Threads, Base clock 2.45 GHz, Up to 3.5GHz
  - 512 GB DDR4 memory
  - 1x AMD Instinct MI210 GPU
    - 64 GB High-Bandwidth Memory (HBM2E)
    - 1.6 TB/s peak bandwidth
    - 26.5 TFLOPS double-precision peak for modeling & simulation
    - 1 Graphic Compute Die (GCD)
  - 1x HPE Slingshot Interconnect 200 GbE NICs
  - 2x NVMe devices, 2.9 TB each (planned for /tmp)

- The CPU and GPU are from the same families as used on the compute nodes to provide a build environment similar to the run environment on the compute nodes

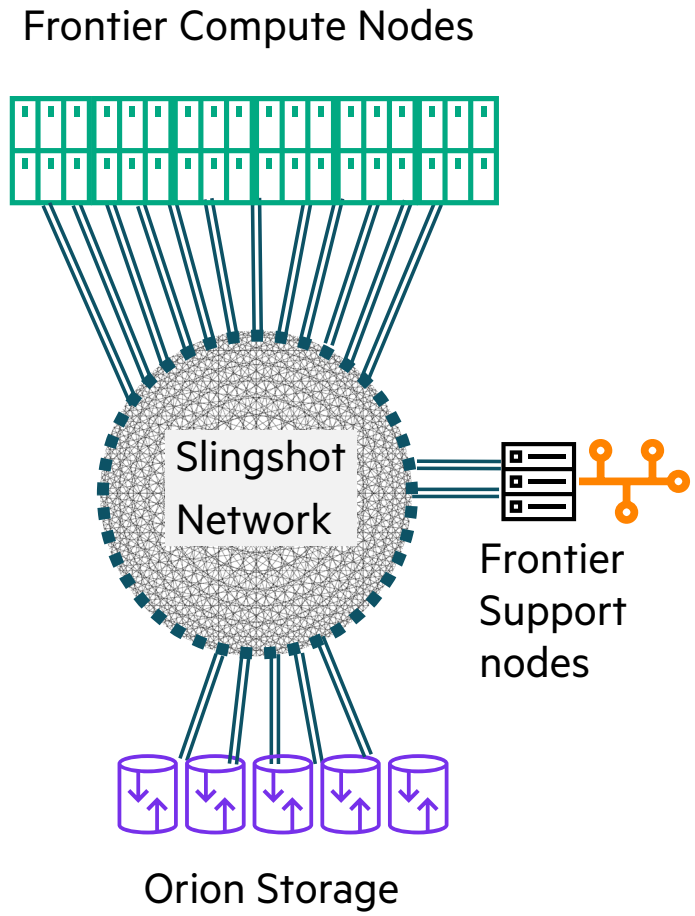


Orion - Lustre Storage for Frontier

- Built on the HPE Cray ClusterStor E1000 family of storage products
- Attached to the HPE Slingshot high-speed network of Frontier
- Directly accessible from each compute node
- Support nodes can provide access to Orion from other OLCF systems

- Uses open-source Lustre parallel file system
- Provides global single namespace POSIX file system
- Supports both file-per-process (N:N) files and shared files (N:1 or N:M)

- Multi-tier design:
  - Performance tier, NVMe device based: 11.5 PB; Bandwidth 10 TB/s R&W
  - Capacity tier, based on HDDs: 679 PB; Bandwidth 5.5 TB/s R, 4.6 TB/s W
  - Metadata tier, NVMe device based: 10 PB
- Orion has 40 MDS nodes, 450 OSS nodes, and 1,350 OSTs



https://www.olcf.ornl.gov/2021/05/20/olcf-announces-storage-specifications-for-frontier-exascale-system/

OLCF Center-Wide Storage and Frontier

- The Frontier architecture also supports access to other OLCF filesystems

- Frontier nodes mount the center-wide NFS-based filesystems
  - Provides User Home (/ccs/home/…)
  - And Project Home (/ccs/project/…)

- Frontier currently mounts the GPFS Alpine IBM Spectrum Scale™ parallel filesystem *
  - Provides 250 PB of storage capacity (/gpfs/alpine/…)
  - Peak write speed of the filesystem is ~2.5 TB/s

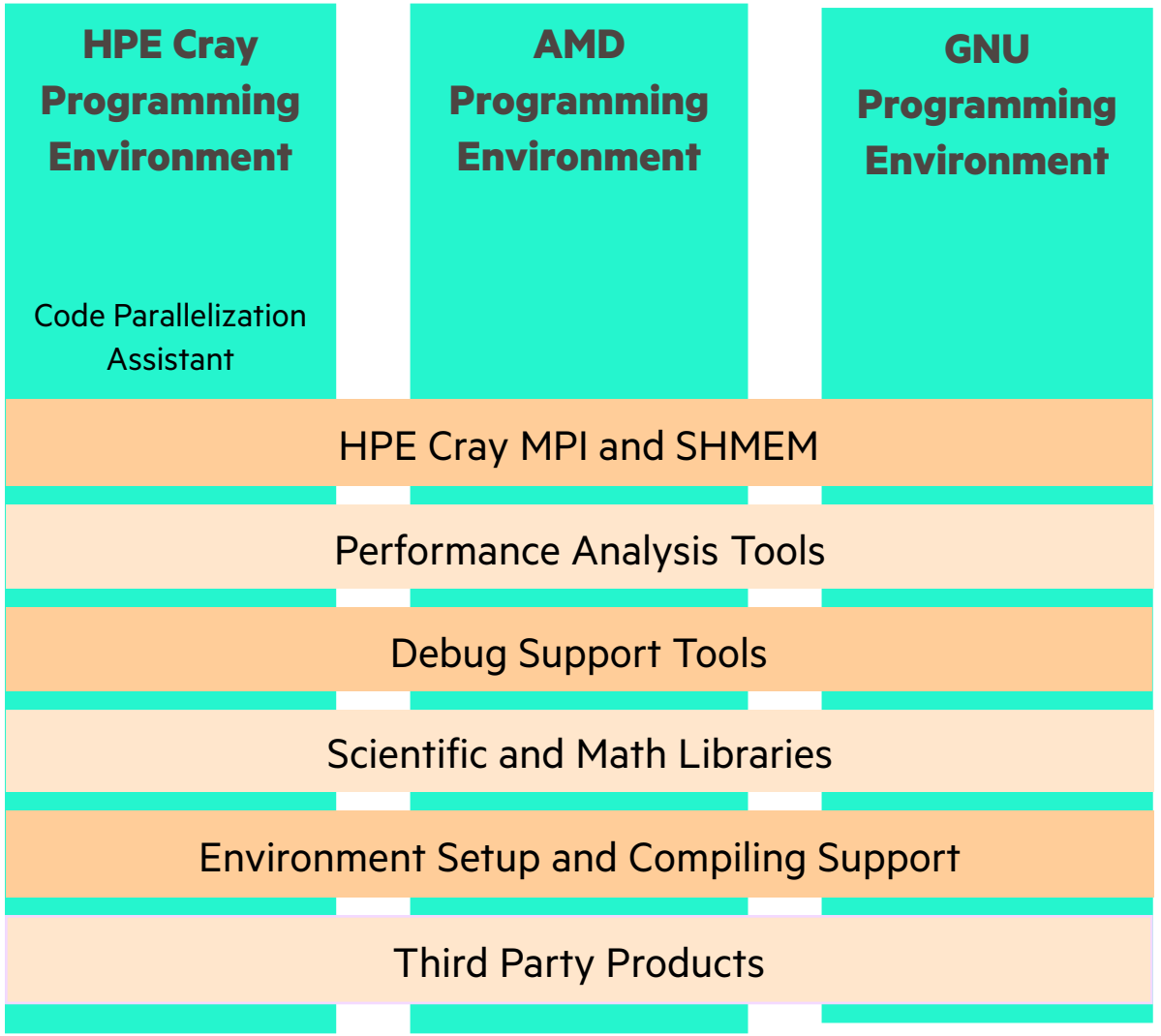
- Note: Frontier does not directly mount the center's High Performance Storage System (HPSS). The Data Transfer Nodes can be used to transfer files between HPSS and Orion

https://docs.olcf.ornl.gov/data/index.html

* When Orion becomes generally available, Frontier is planned to no longer mount Alpine

Application Software Stack for Frontier

- ORNL, HPE, and AMD are working together to deliver a full software stack for Frontier
  - Provides compiler and library choice, performance, and programmability
  - Includes:
    - Multiple programming environments
    - Performance and correctness tools
    - Optimizations such as:
      - MPI GPU-to-GPU data movement
      - libsci_acc
      - DL Plugin
    - Compiler interoperability
  - HPE and AMD continue to enhance the Frontier software stack
- Frontier will get updated versions of the software as they become available



Thank you

glenski@hpe.com