OAK RIDGE National Laboratory

Odo Architecture Overview

Subil Abraham

HPC Engineer – User Assistance

July 8, 2024

U.S. DEPARTMENT OF ENERGY

Frontier System Overview

- HPE Cray EX Supercomputer architecture
- 74 cabinets, 128 nodes per cabinet (9408 nodes)
- 3rd Gen AMD EPYC 64-core CPU
- 4 AMD Instinct MI250X GPUs
- HPE Slingshot interconnect
- Peak 1.206 Exaflops on HPL benchmark
- Cray, AMD, and GNU software stacks
- Orion - 679 PB multi-tier Lustre filesystem
- NFS storage (/ccs/home, /ccs/proj)

Frontier System Overview

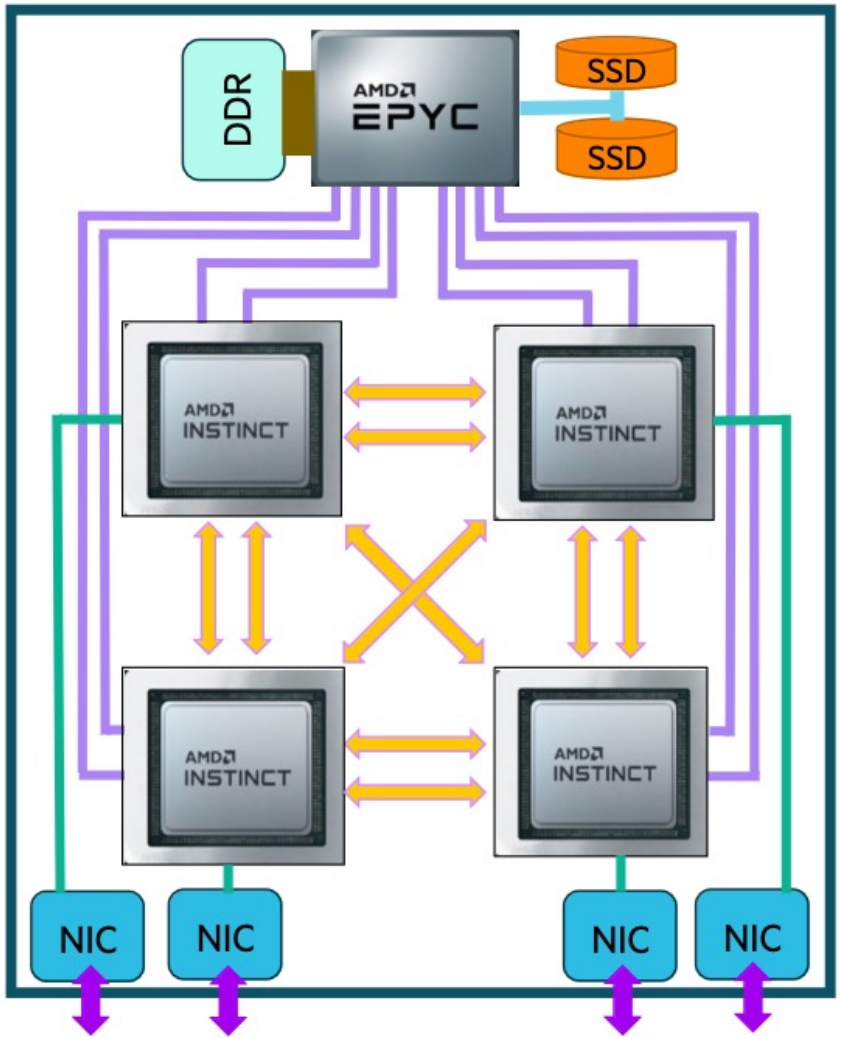
- HPE Cray EX Supercomputer architecture
- 74 cabinets, 128 nodes per cabinet (9408 nodes)
- [illegible] CPU

Odo is 30 Frontier nodes, and uses the GPFS filesystem (/gpfs/alpine2), and NFS on the Open Enclave for home areas (/ccsopen/home, /ccsopen/proj)

- Orion - 679 PB multi-tier Lustre filesystem
- NFS storage (/ccs/home, /ccs/proj)

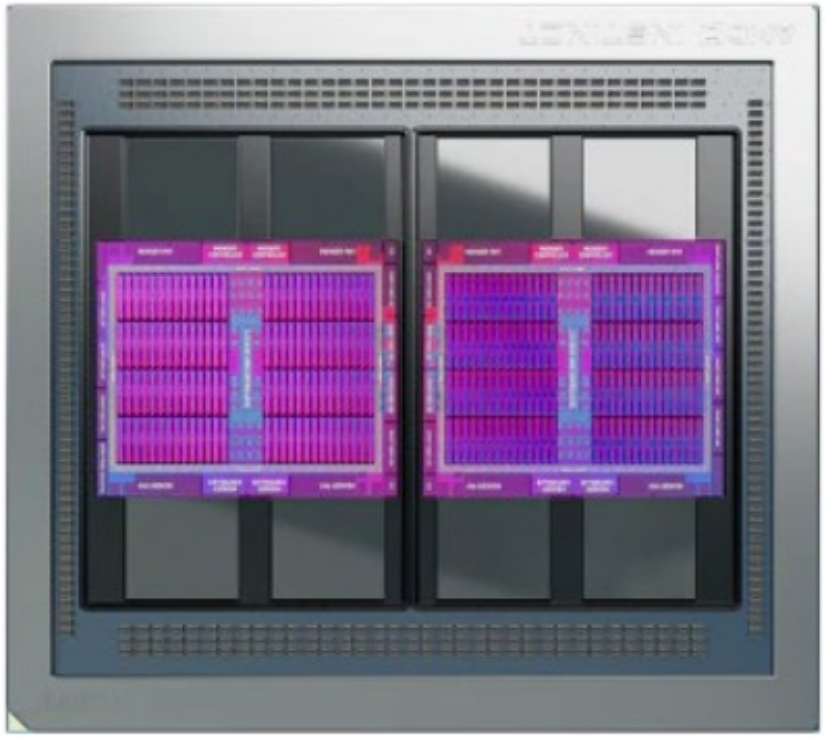
Frontier Compute Node Design

- 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
- 4x HPE Slingshot Interconnect 200 GbE NICs
  - Provides 100 GB/s to other nodes, 25 GB/s per port

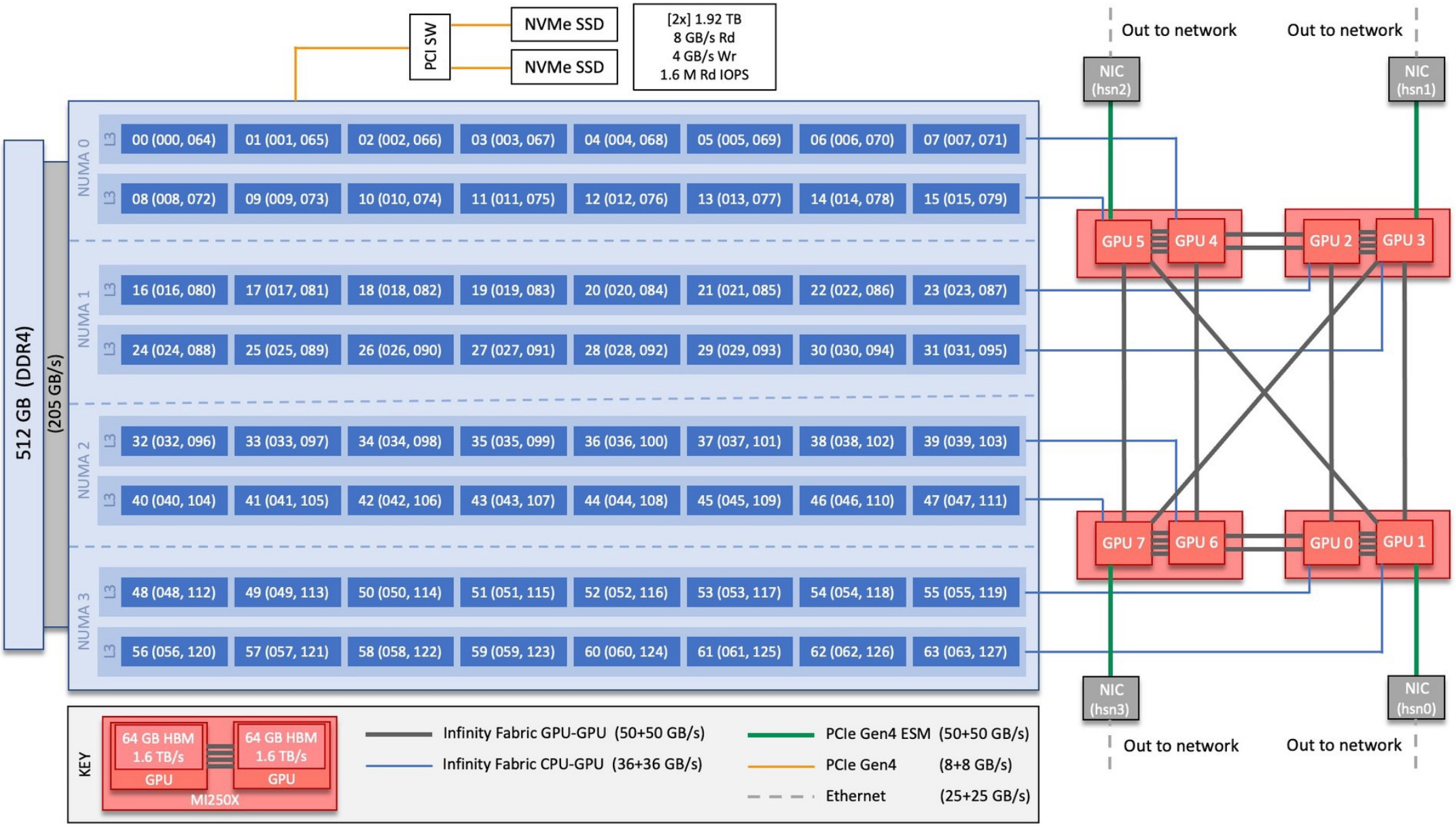


MI250X Accelerators

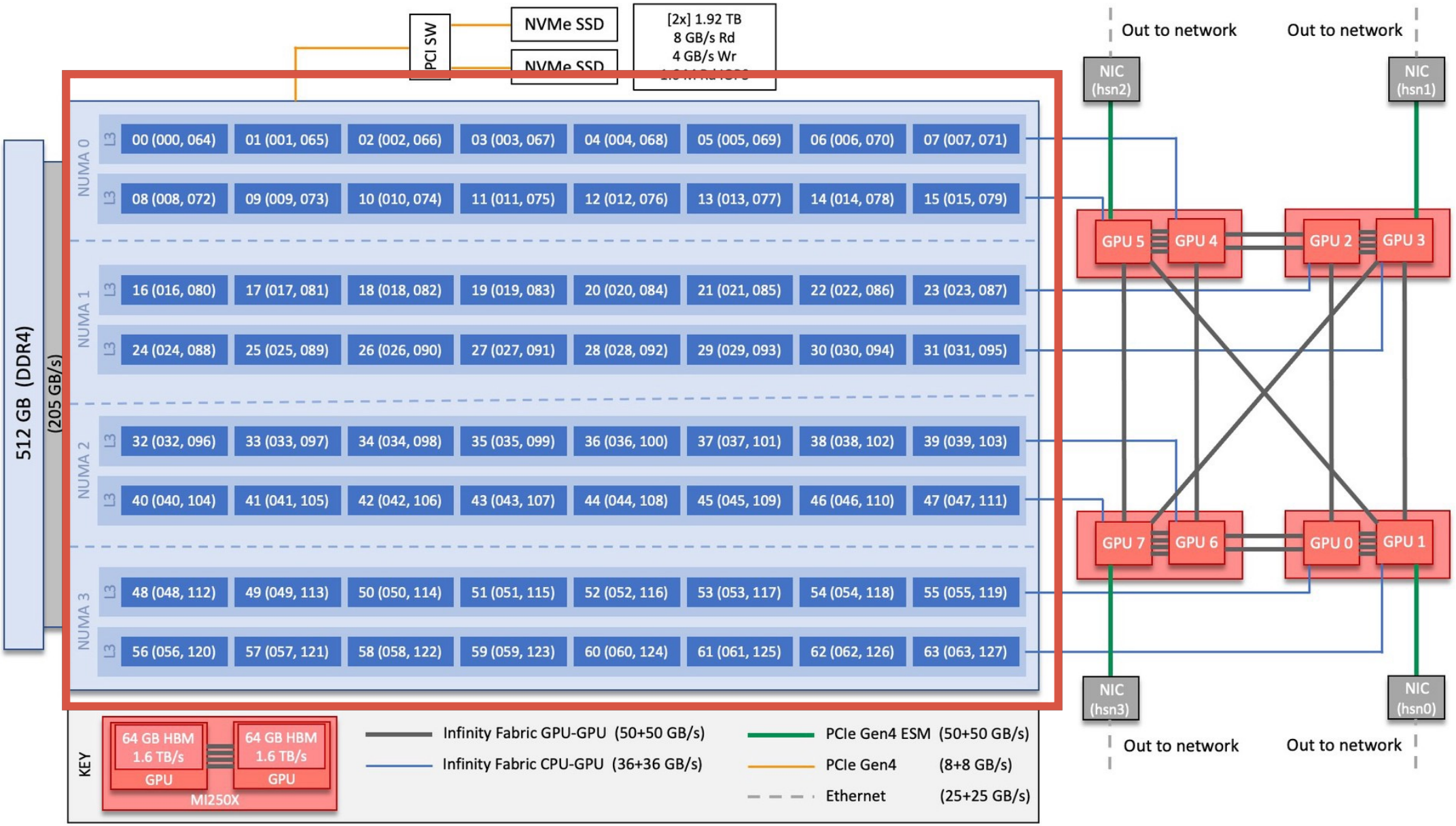
- The AMD MI250X has two Graphic Compute Dies (GCDs) per module
- This gives a total of 8 GCDs per node
- The 8 GCDs show as 8 separate GPUs to the OS, Slurm, and ROCm
- Generally easier to refer to the 8 GCDs as GPUs and arrange programs to run on nodes with 8 GPUs
- Each Node then has 8 GPUs with the 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



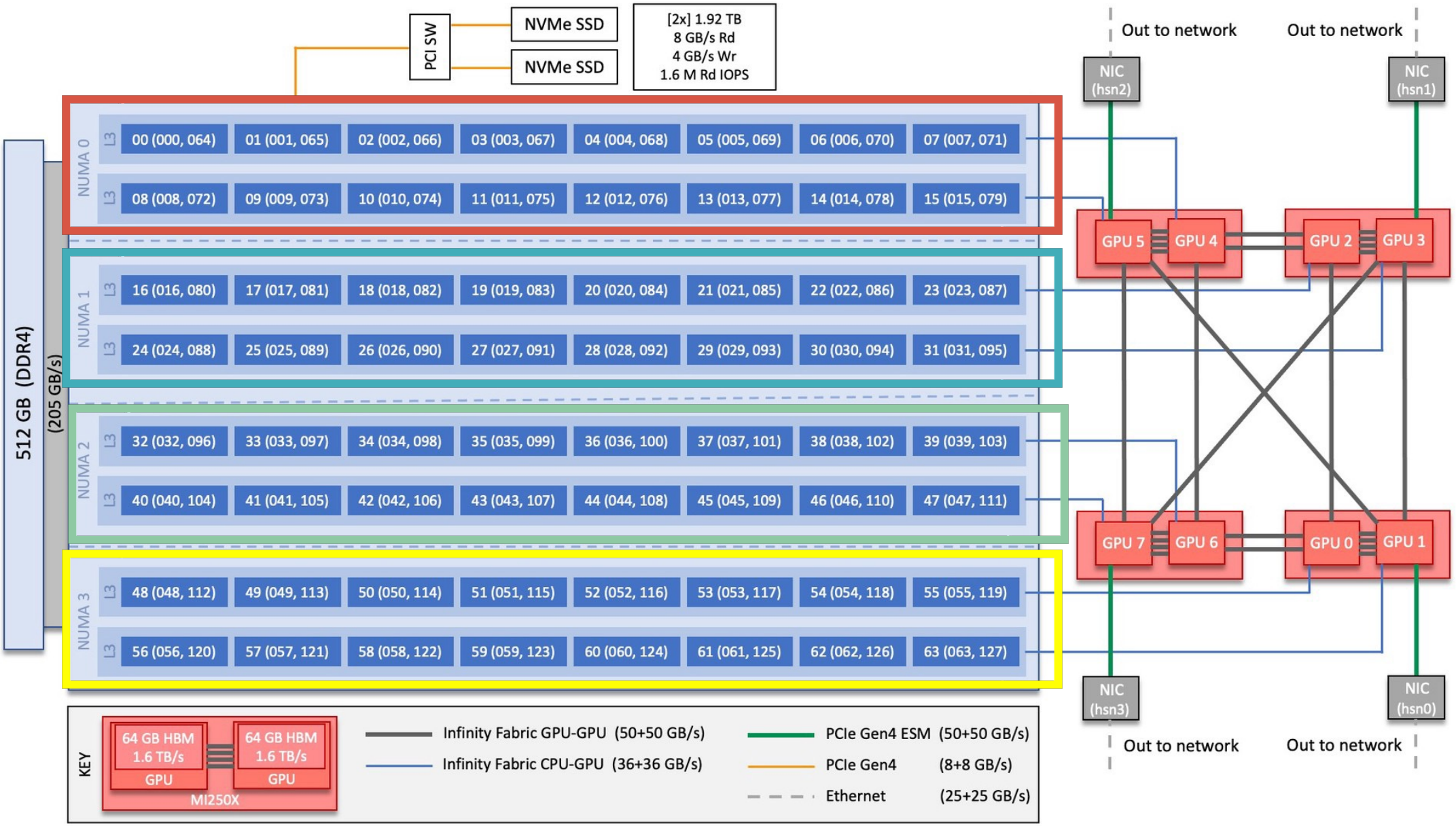
Frontier Compute Node Diagram



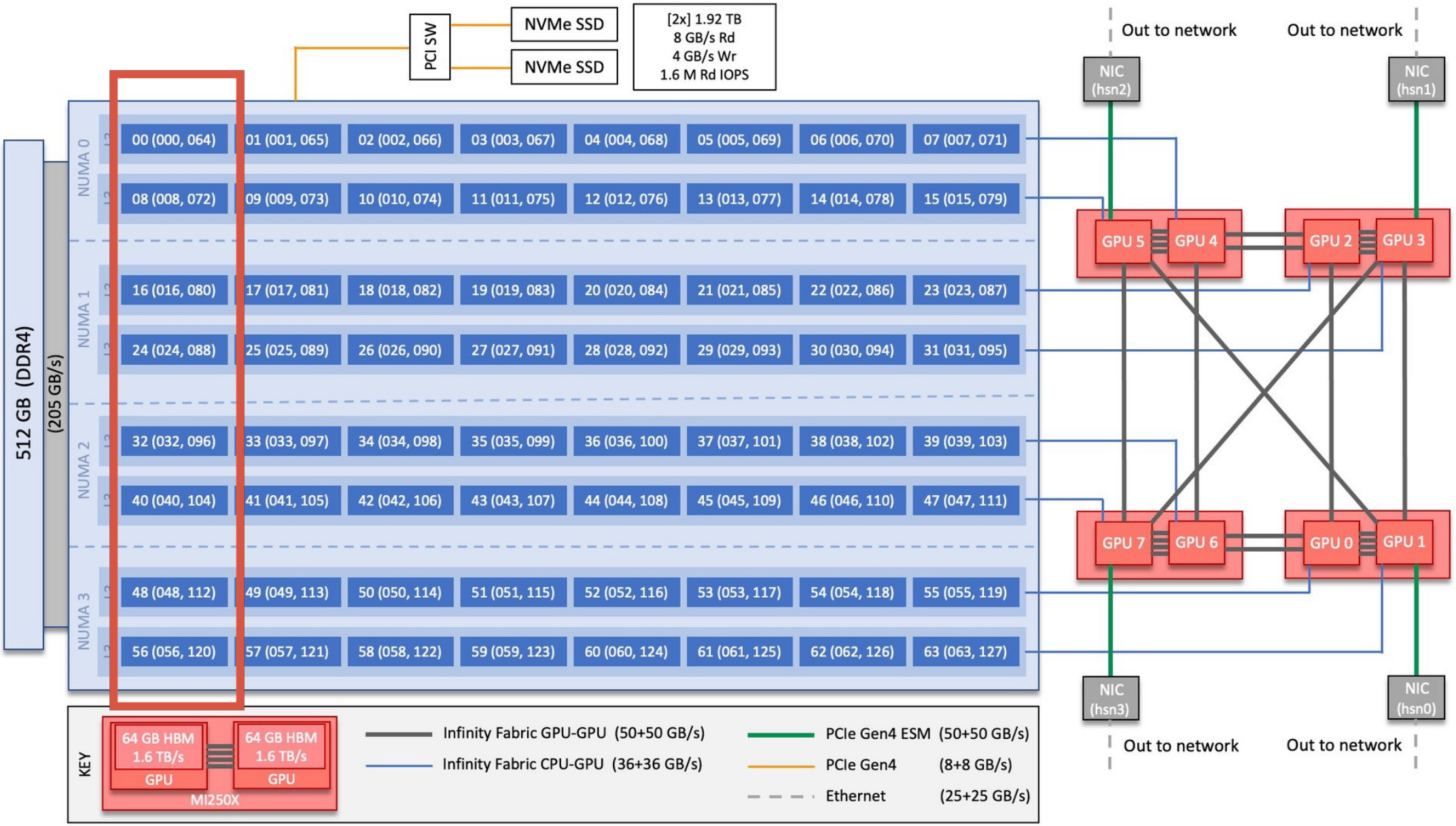
Frontier Compute Node Diagram



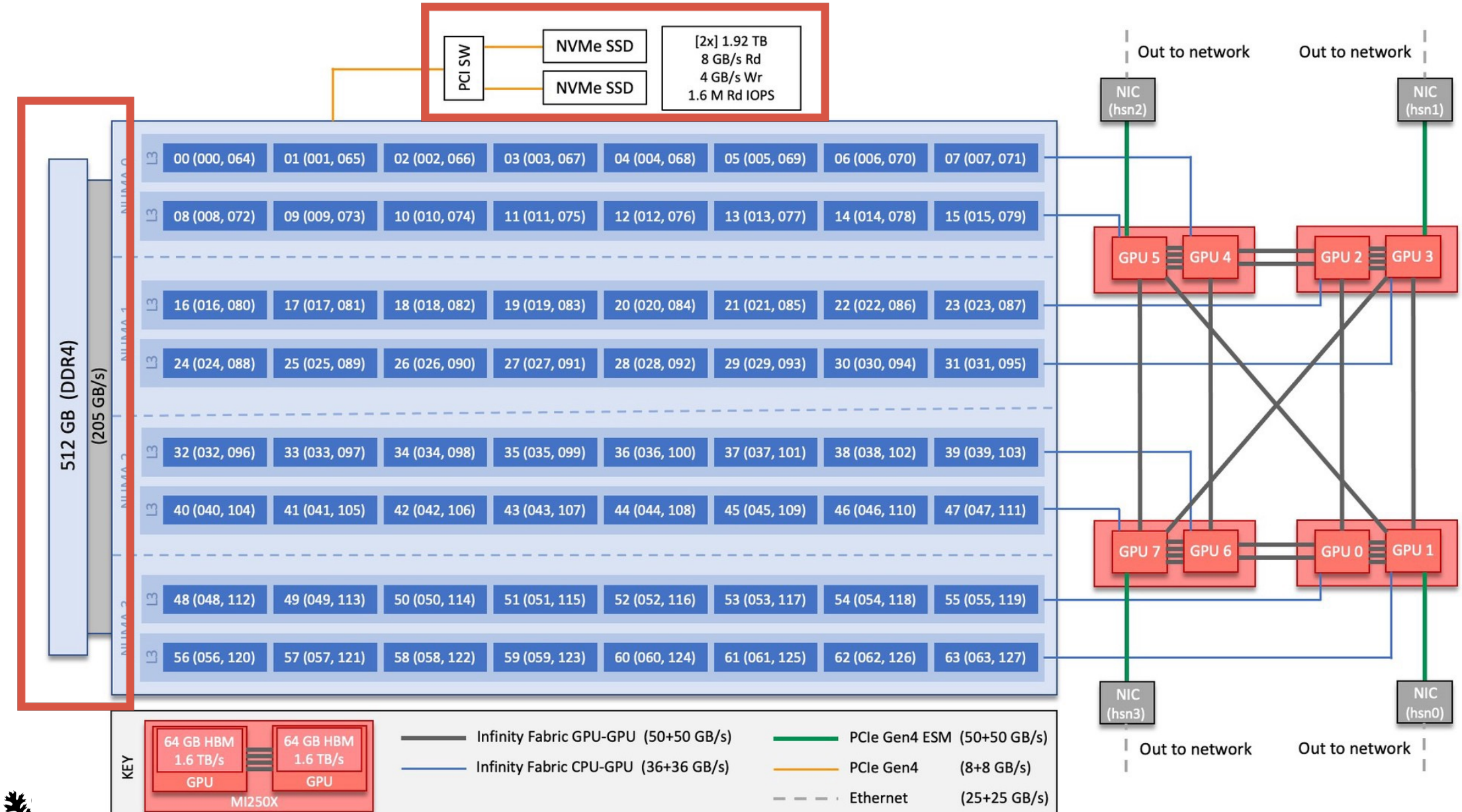
Frontier Compute Node Diagram



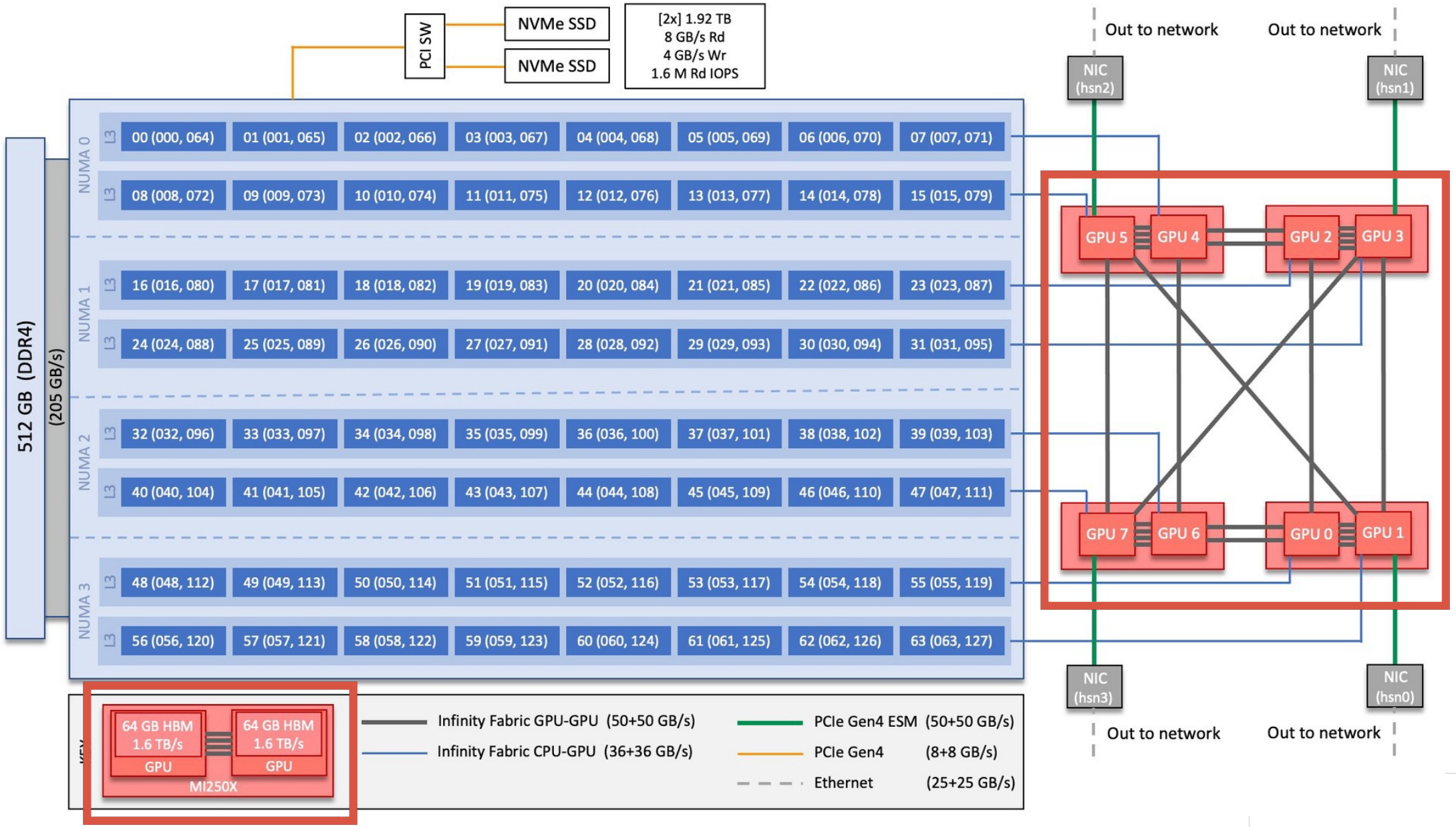
Frontier Compute Node Diagram



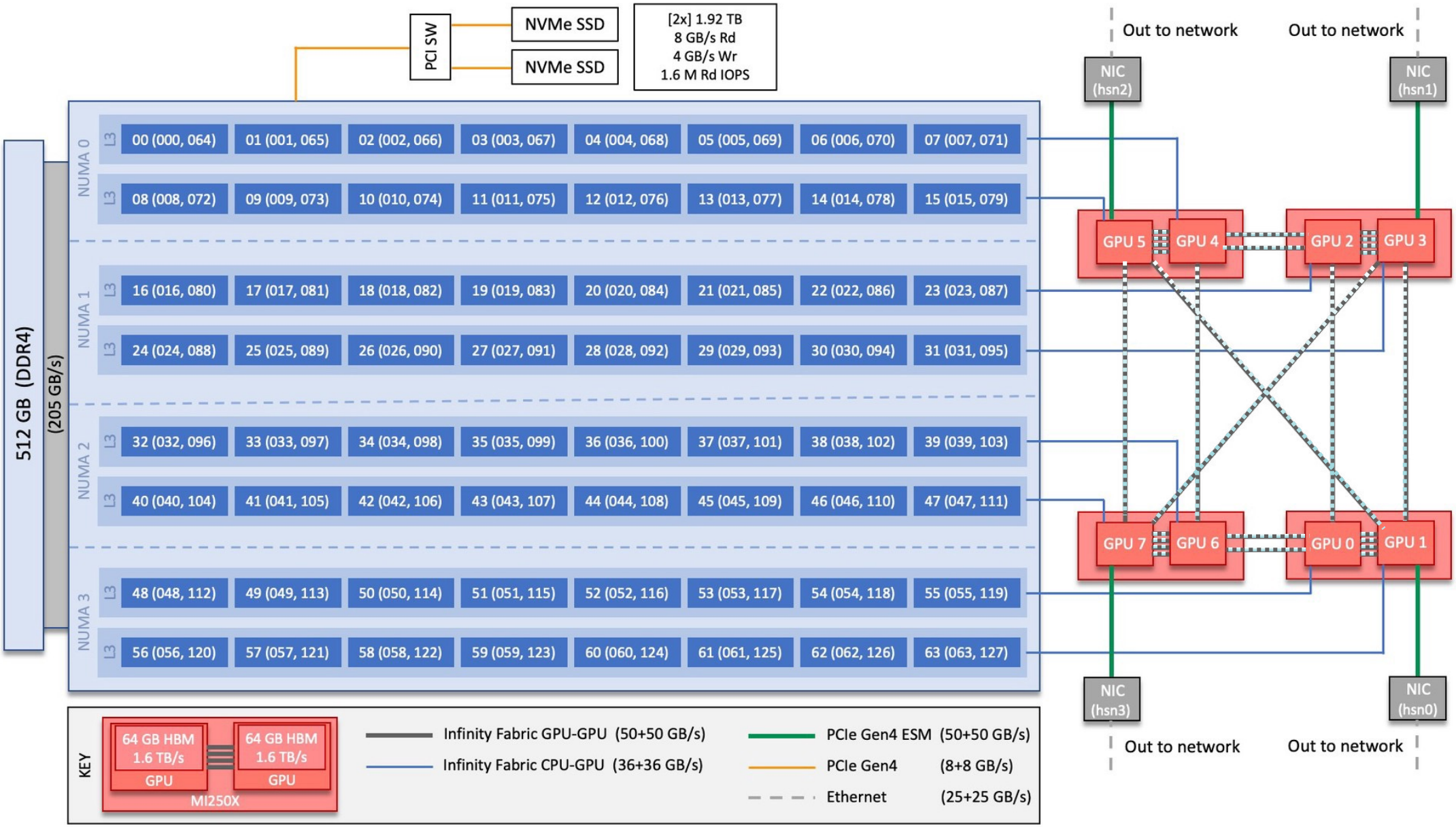
Frontier Compute Node Diagram



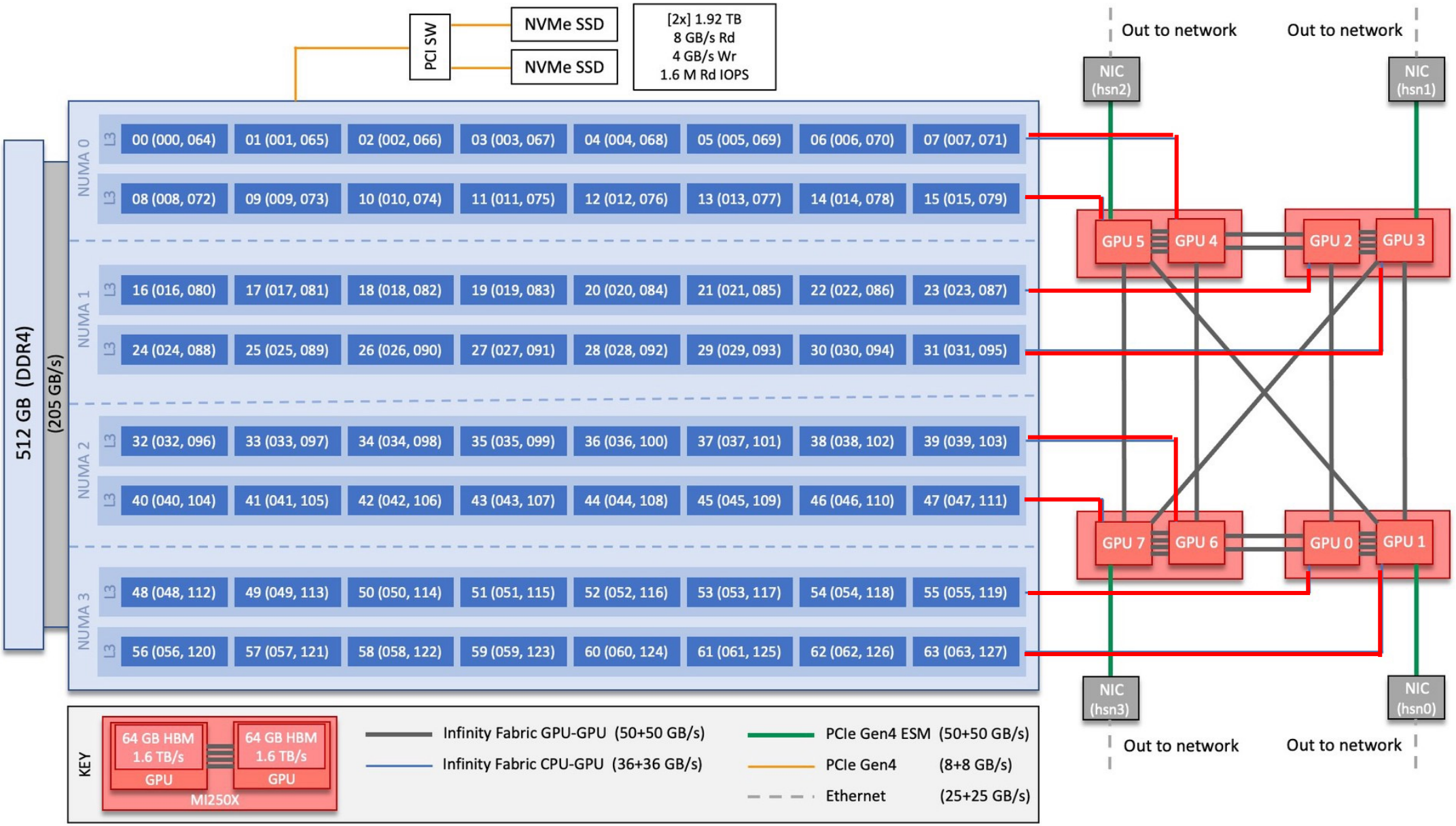
Frontier Compute Node Diagram



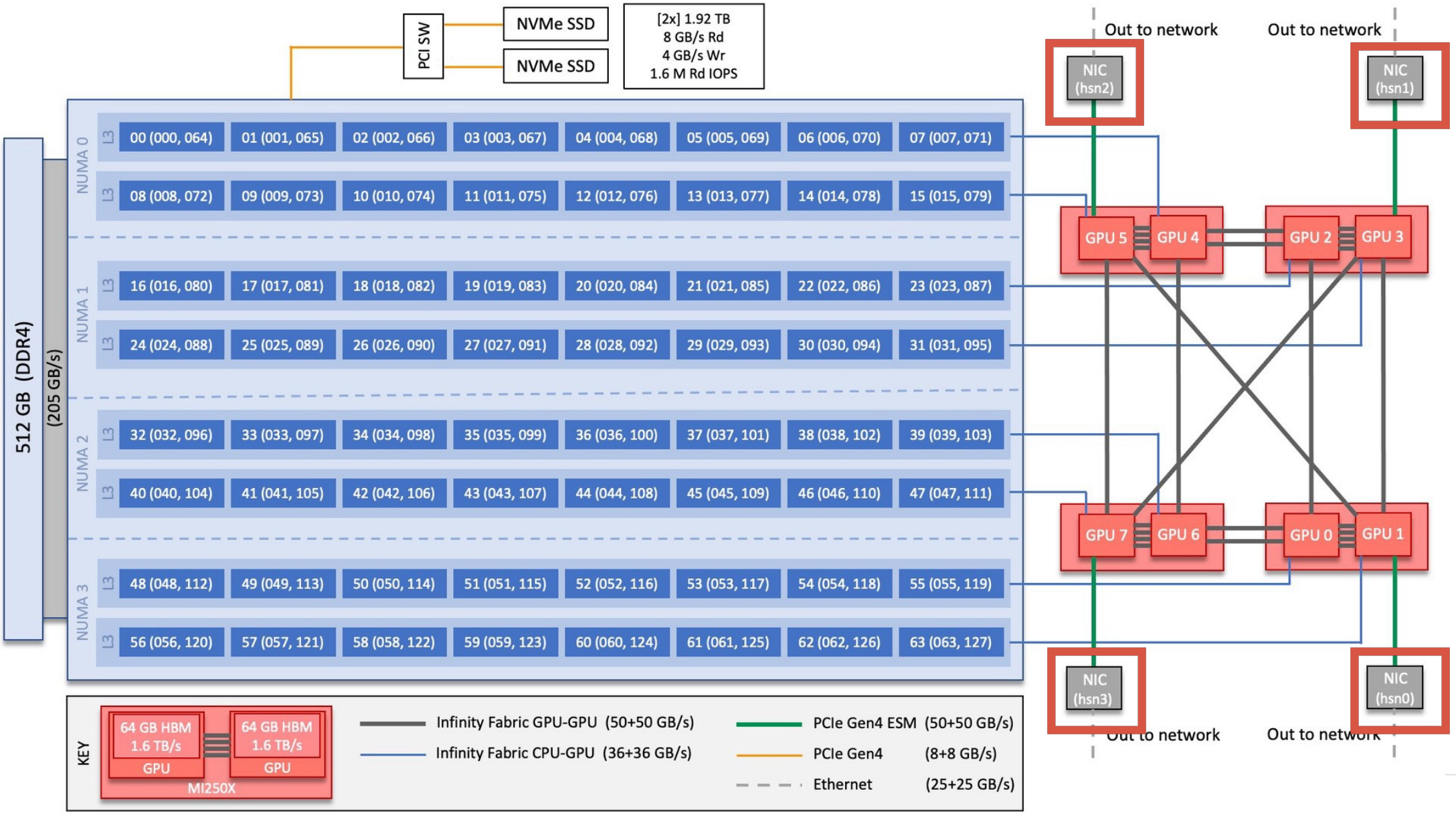
Frontier Compute Node Diagram



Frontier Compute Node Diagram



Frontier Compute Node Diagram



For everything we talked about!

Documentation:
docs.olcf.ornl.gov

Joe Glenski's 30 minute talk:
https://vimeo.com/840551316