Frontier's Architecture

Scott Atchley

Preparing For Frontier Training Series

July 12, 2022

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

Agenda

- OLCF Leadership Systems
- Frontier Node Overview
- Frontier's Interconnect

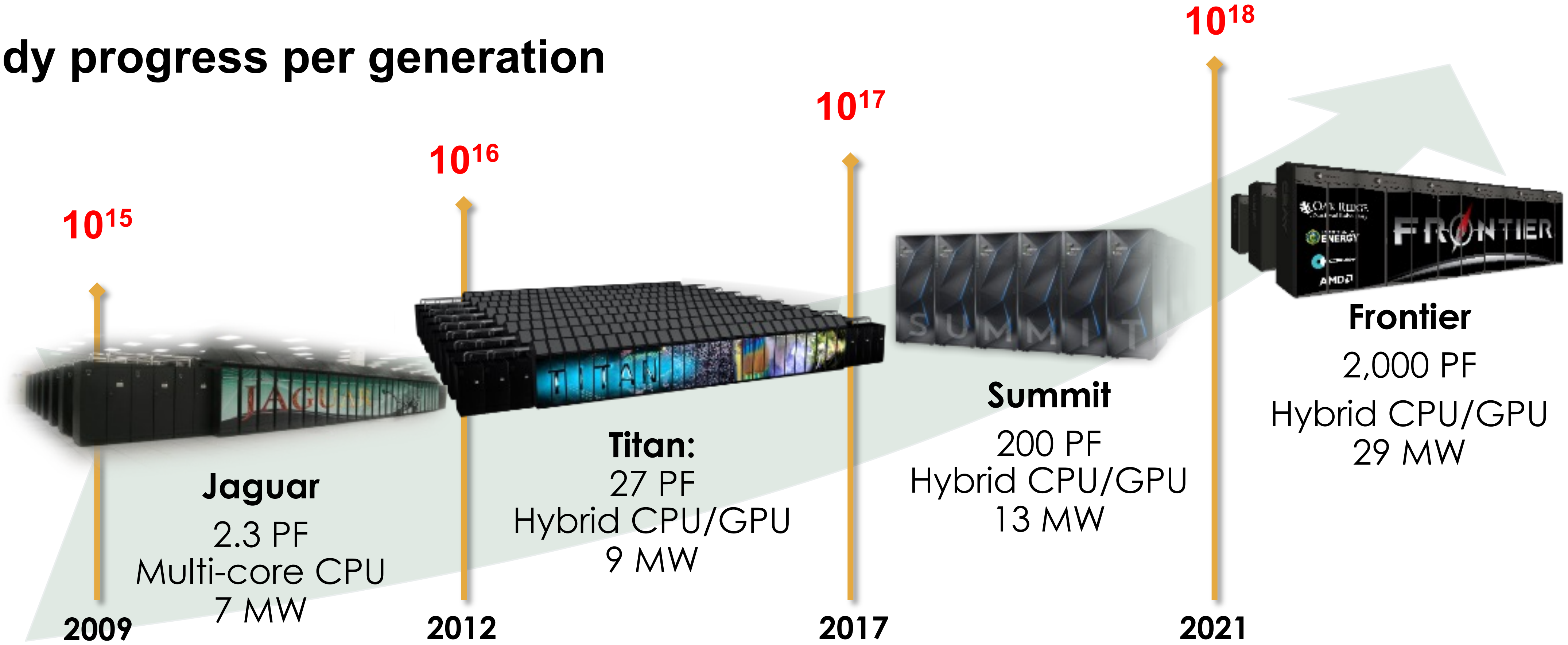
OLCF Leadership Systems

From Petascale to Exascale

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

Vision: Deliver transforming discoveries in energy technologies, materials, biology, environment, health, etc.

Steady progress per generation



10^{15}

Jaguar

2.3 PF

Multi-core CPU

7 MW

2009

10^{16}

Titan:

27 PF

Hybrid CPU/GPU

9 MW

2012

10^{17}

Summit

200 PF

Hybrid CPU/GPU

13 MW

2017

10^{18}

Frontier

2,000 PF

Hybrid CPU/GPU

29 MW

2021

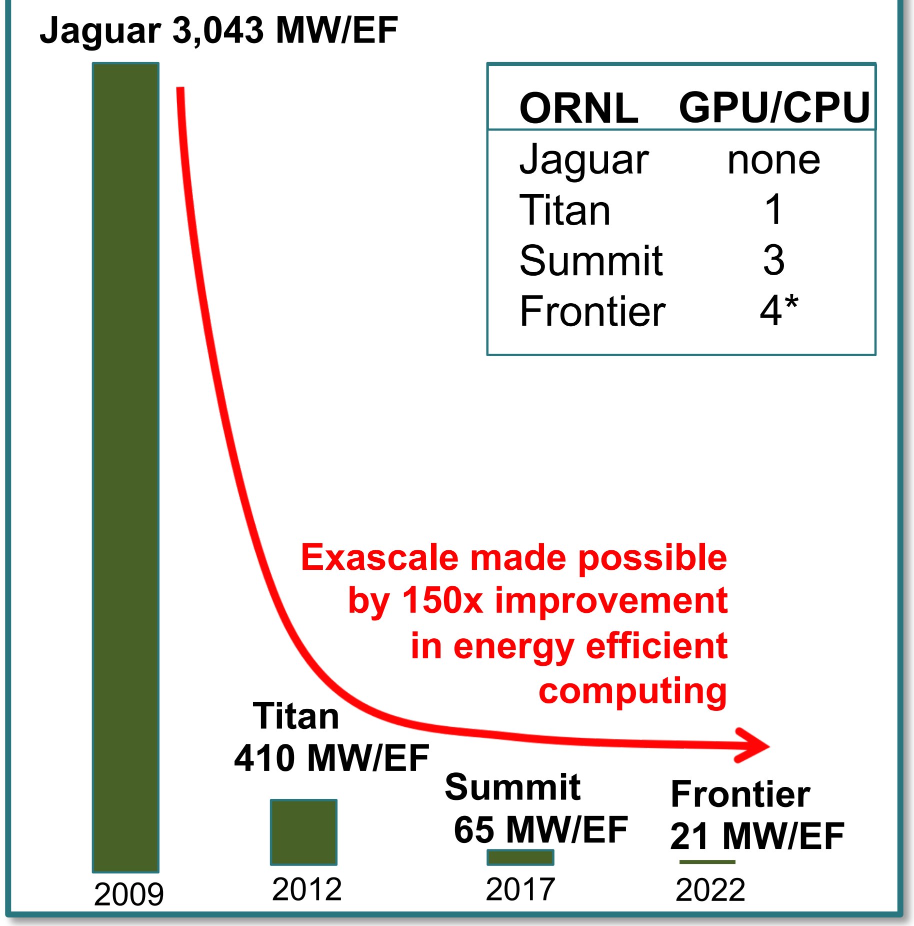
Energy Efficiency - One of the key Exascale challenges

Since 2008, one of the biggest concerns with reaching Exascale has been energy consumption

- **ORNL pioneered GPU use in supercomputing** beginning in 2012 with Titan thru today with Frontier. Significant part of energy efficiency improvements.
- **DOE *Forward vendor investments** in energy efficiency (2012-2020) further reduced the power consumption of computing chips (CPUs and GPUs).
- **150x reduction in energy per FLOPS** from Jaguar to Frontier at ORNL
- ORNL achieves additional energy savings from using warm water cooling in Frontier (32 C).
  ORNL Data Center PUE= 1.03

Frontier first US Exascale computer

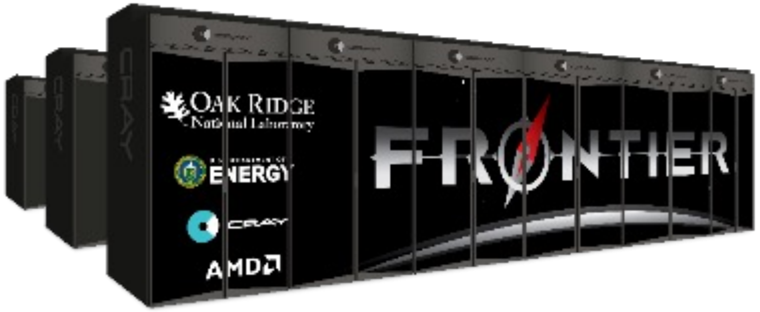
Multiple GPU per CPU drove energy efficiency



ORNL	GPU/CPU
Jaguar	none
Titan	1
Summit	3
Frontier	4*

Frontier Overview Built by HPE Powered by AMD

Extraordinary Engineering

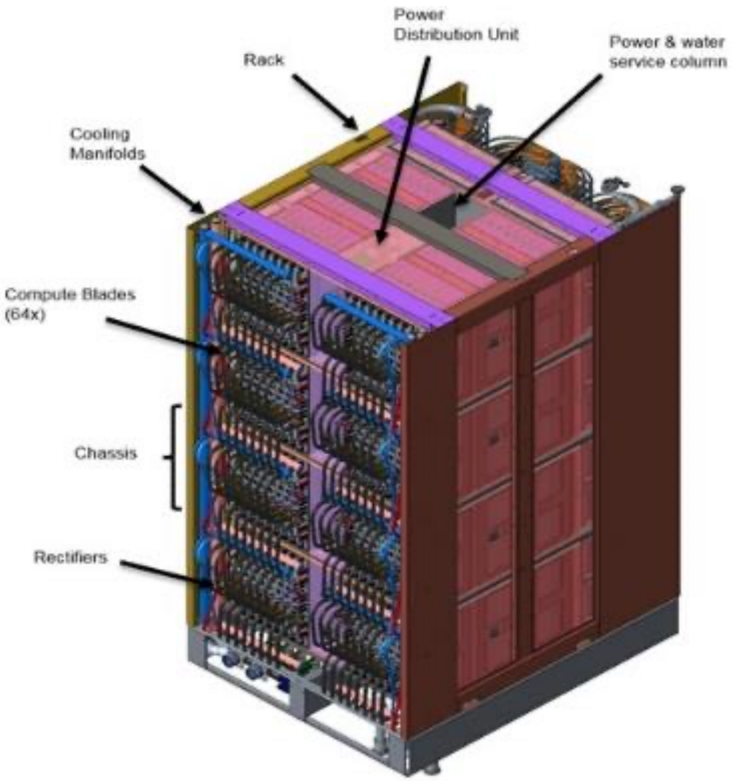


System

- 2.0 EF Peak DP FLOPS
- 74 compute racks
- 29 MW Power Consumption
- 9,408 nodes
- 9.2 PiB memory
  (4.6 PiB HBM, 4.6 PiB DDR4)
- Cray Slingshot network with dragonfly topology
- 37 PB Node Local Storage
- 716 PB Center-wide storage
- 4,000 ft^2 footprint

Olympus rack

- 128 AMD nodes
- 8,000 lbs
- Supports 400 KW



AMD node

- 1 AMD “Trento” CPU
- 4 AMD MI250X GPUs
- 512 GiB DDR4 memory on CPU
- 512 GiB HBM2e total per node
  (128 GiB HBM per GPU)
- Coherent memory across the node
- 4 TB NVM
- GPUs & CPU fully connected with AMD Infinity Fabric
- 4 Cassini NICs, 100 GB/s network BW

Compute blade

- 2 AMD nodes

All water cooled, even DIMMS and NICs

One more word on power efficiency

- One cabinet of Frontier has a 10% higher HPL than all of Titan
  - While only using 309 kW compared to the Titan’s 7 MW

>

One Cabinet
24 ft^2

200 Cabinets
~4,500 ft^2

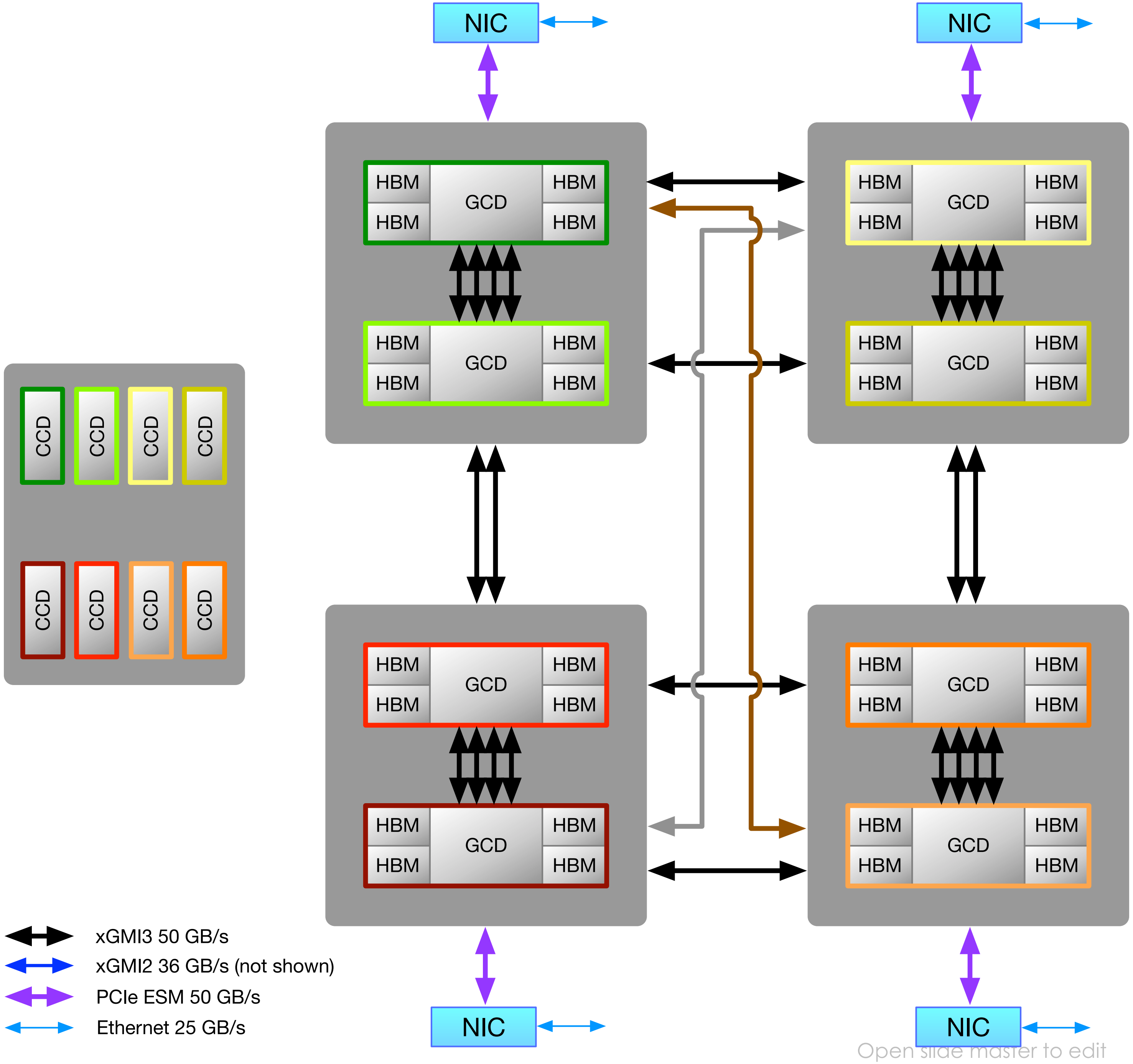
OLCF Systems by the numbers

System	Titan (2012)	Summit (2017)	Frontier (2021)
Peak	27 PF	200 PF	2.0 EF
# nodes	18,688	4,608	9,408
Node	1 AMD Opteron CPU 1 NVIDIA Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 AMD EPYC “Trento” CPU 4 AMD Instinct MI250X GPUs
Memory	0.6 PB DDR3 + 0.1 PB GDDR	2.4 PB DDR4 + 0.4 HBM + 7.4 PB On-node storage	4.6 PB DDR4 + **4.6 PB HBM2e** + 36 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™	695 PB HDD+11 PB Flash Performance Tier, 9.4 TB/s and 10 PB Metadata Flash Lustre
Power	9 MW	13 MW	29 MW

Frontier Node Overview

Bard Peak Node

- Trento has 8 CCDs
- Each MI250X has two GCDs
  - Each GCD appears as a GPU to the user
  - Each node has **8 GPUs**
- One GCD per CCD
  - xGMI2 links each pair
- 1 NIC attached to each MI250X
  - HBM resident data avoids slower CPU link



OLCF Systems by the numbers revisited

System	Titan (2012)	Summit (2017)	Frontier (2021)
CPU:GPU	1:1	1:3	1:8
CPU Mem BW	50 GB/s	170 GB/s per CPU	205 GB/s
GPU Mem BW	1x 250 GB/s 250 GB/s Total	3x 900 GB/s 2,700 GB/s Total	8x 1,635 GB/s 13,080 GB/s Total
Interconnect BW	1x 6 GB/s 6 GB/s Total	3x 50 GB/s 150 GB/s Total	8x 36 GB/s 288 GB/s Total
Fast-to-Slow Memory Ratio	~~5:1 GPU:CPU~~ 42:1 GPU:CPU limited by PCIe	16:1 not limited by NVLink	**64:1** not limited by xGMI-2

- Titan's ratio was too slow to effectively use the host memory
- **Frontier's ratio is much worse**
  - Each Frontier has more than 5x the HBM than a Summit node
  - **Size your application to fit in HBM**
  - The host memory is good for caching data that would be read from/written to the file system

Frontier's Interconnect

OLCF System Interconnects

Interconnect
Cray SeaStar

Node Injection
8 GB/s

Interface
Portals-3

Topology
3D Torus

Interconnect
Cray Gemini

Node Injection
6.4 GB/s

Interface
uGNI

Topology
3D Torus

Interconnect
Mellanox EDR IB

Node Injection
2x 12.5 GB/s

Interface
Verbs

Topology
Clos
(non-blocking
fat-tree)

180+ miles of cables

Interconnect
HPE Slingshot

Node Injection
4x 25 GB/s

Interface
Libfabric/OFI

Topology
Dragonfly

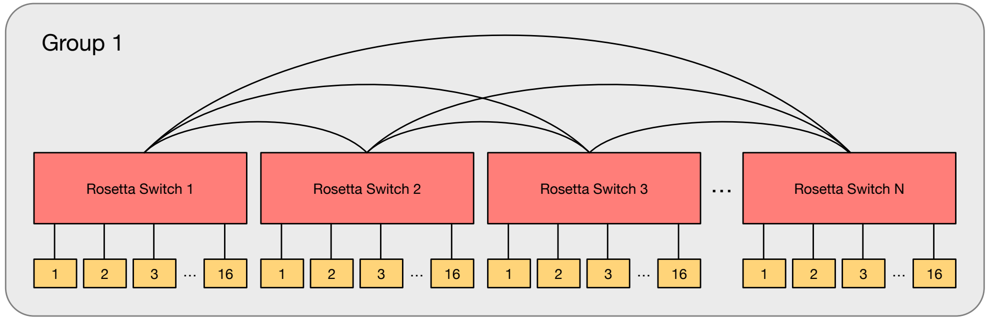
90+ miles of cables

What is Slingshot?

- HPC Ethernet Protocol
  - A superset of Ethernet
  - Optimizes packet headers, reduces padding and interframe gap
  - Negotiated between switch and NIC after link training
    - Otherwise falls back to standard Ethernet
- Hardware
  - Rosetta switches
  - Cassini NICs
    - Accessed via OpenFabrics (aka libfabric)
      - FIFOs, tagged messages, RMA, atomics

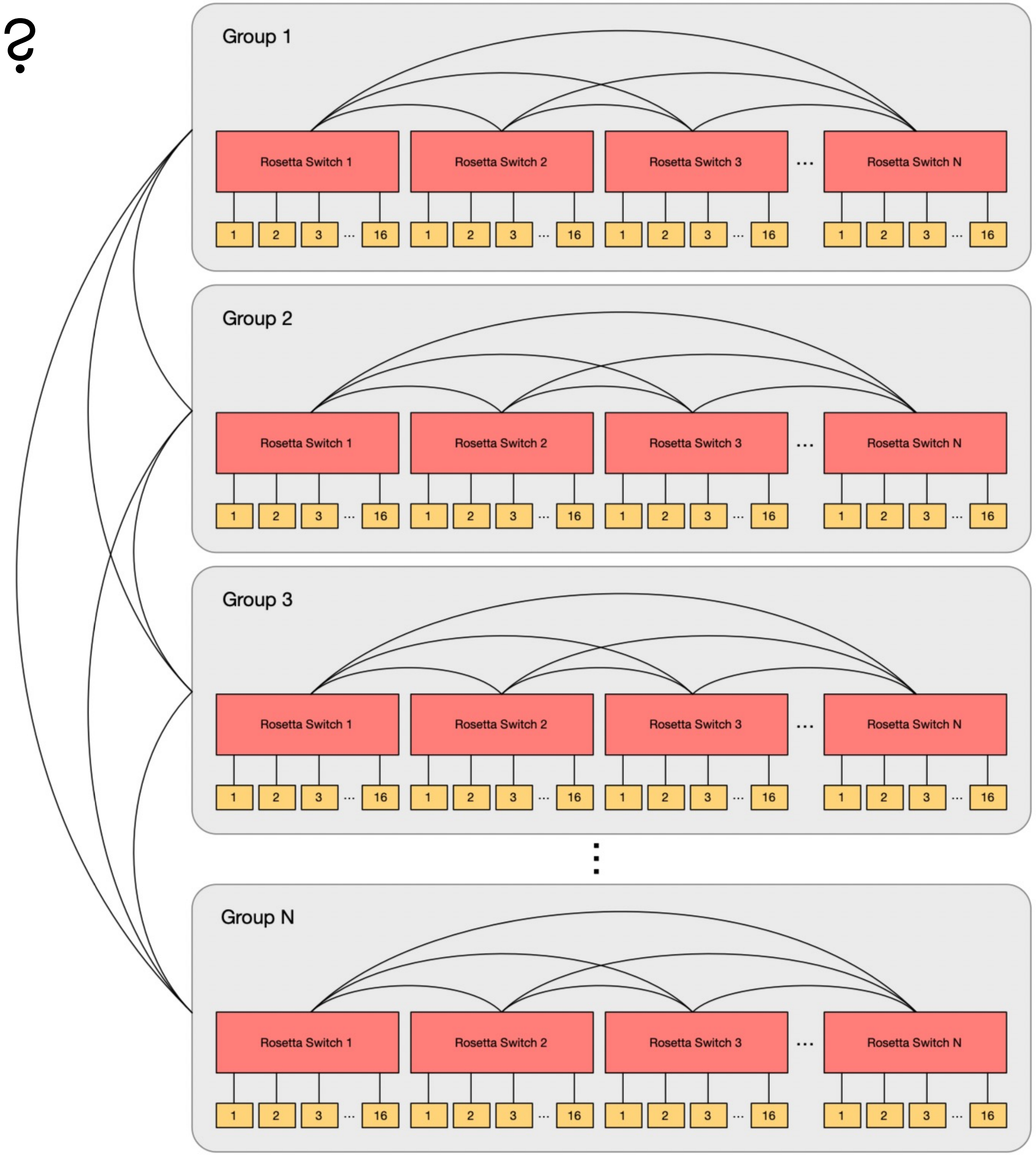
What is a Dragonfly group?

- A group of endpoints connected to switches that are connected all-to-all



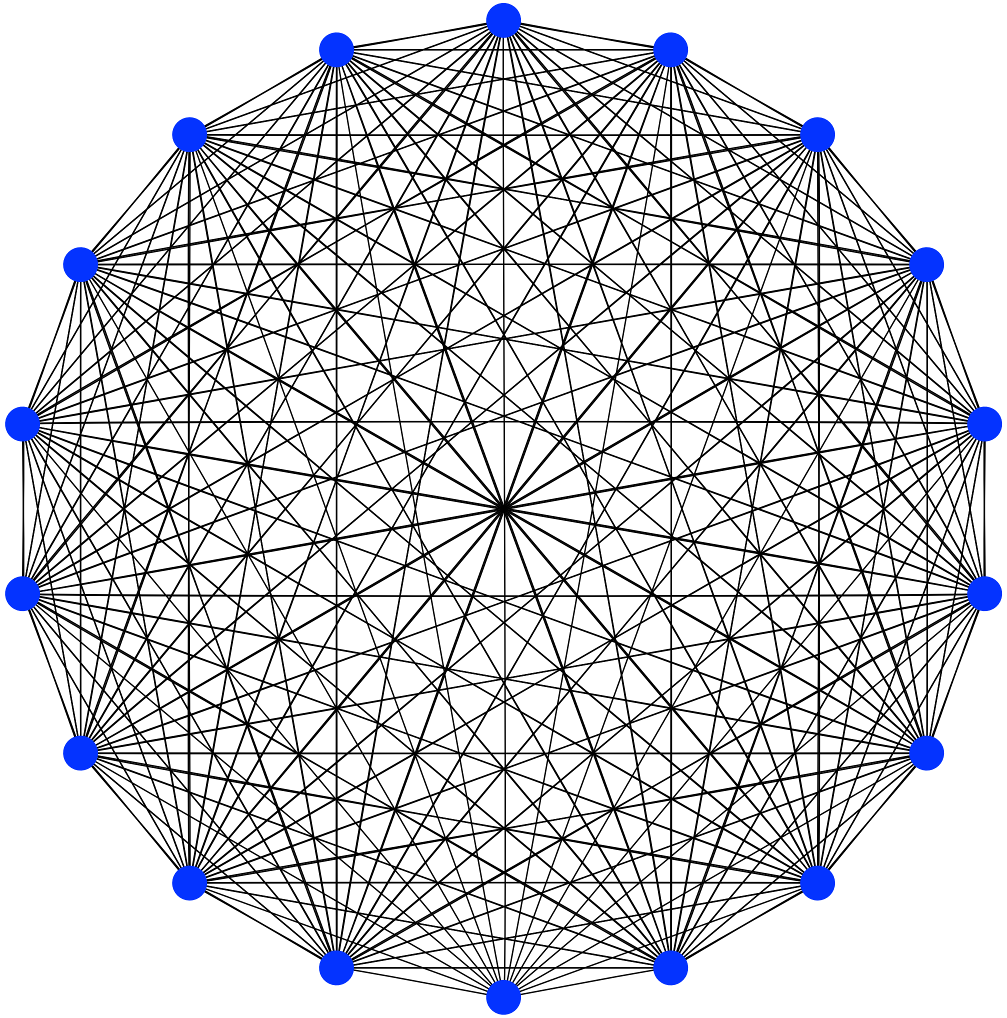
What is a Dragonfly topology?

- A set of groups that are connected all-to-all
  - Every group has one or more links to every other group



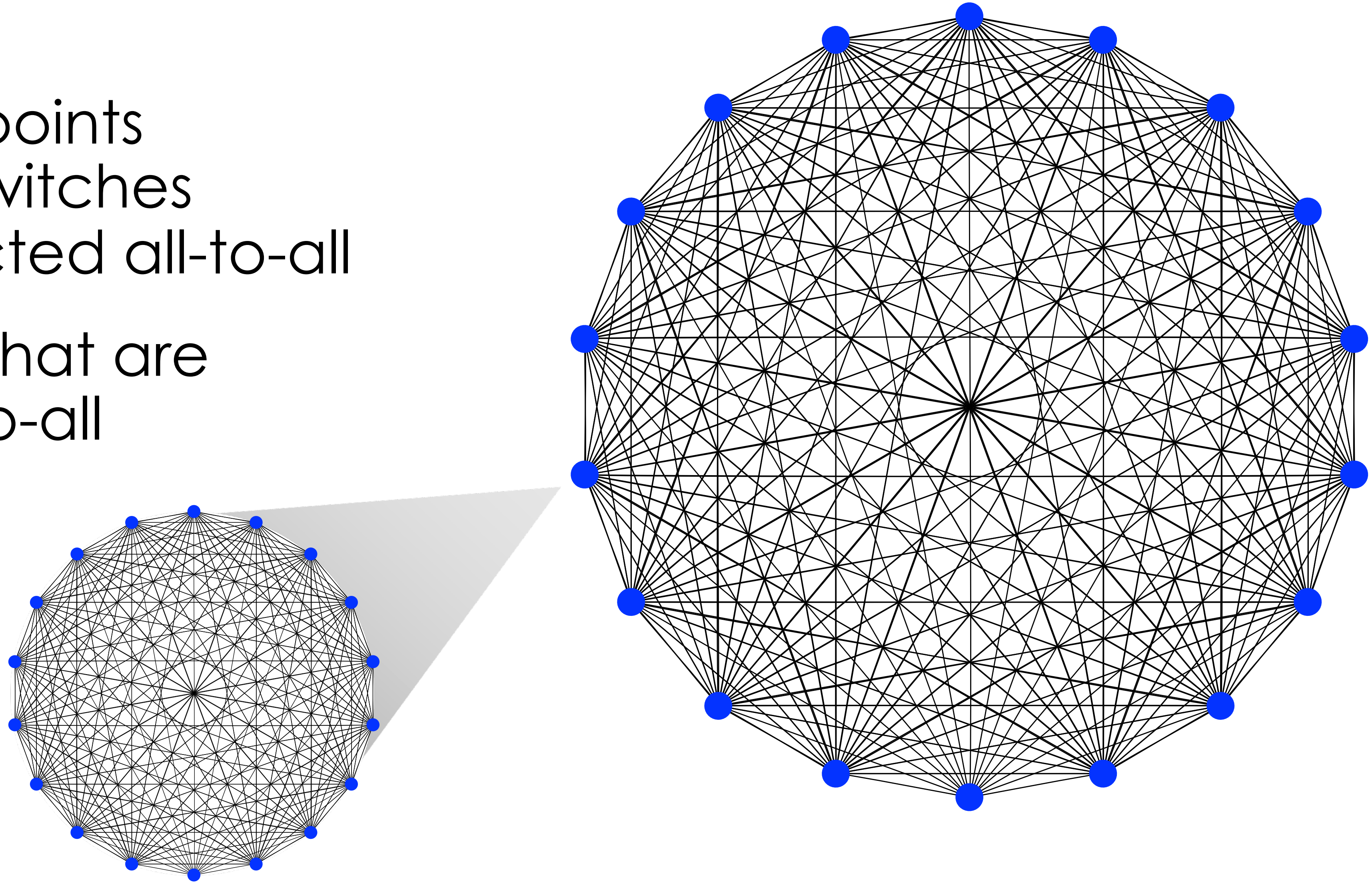
Another view of a Dragonfly Group

- A group of endpoints connected to switches that are connected all-to-all

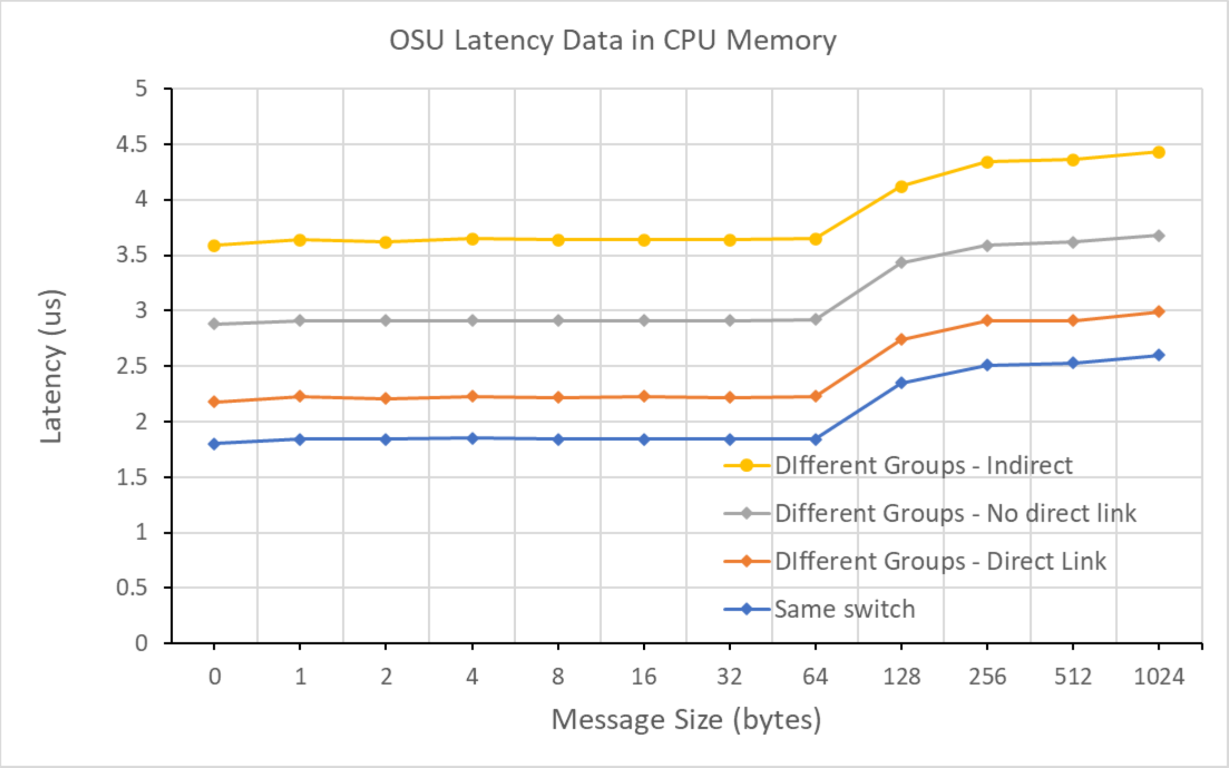


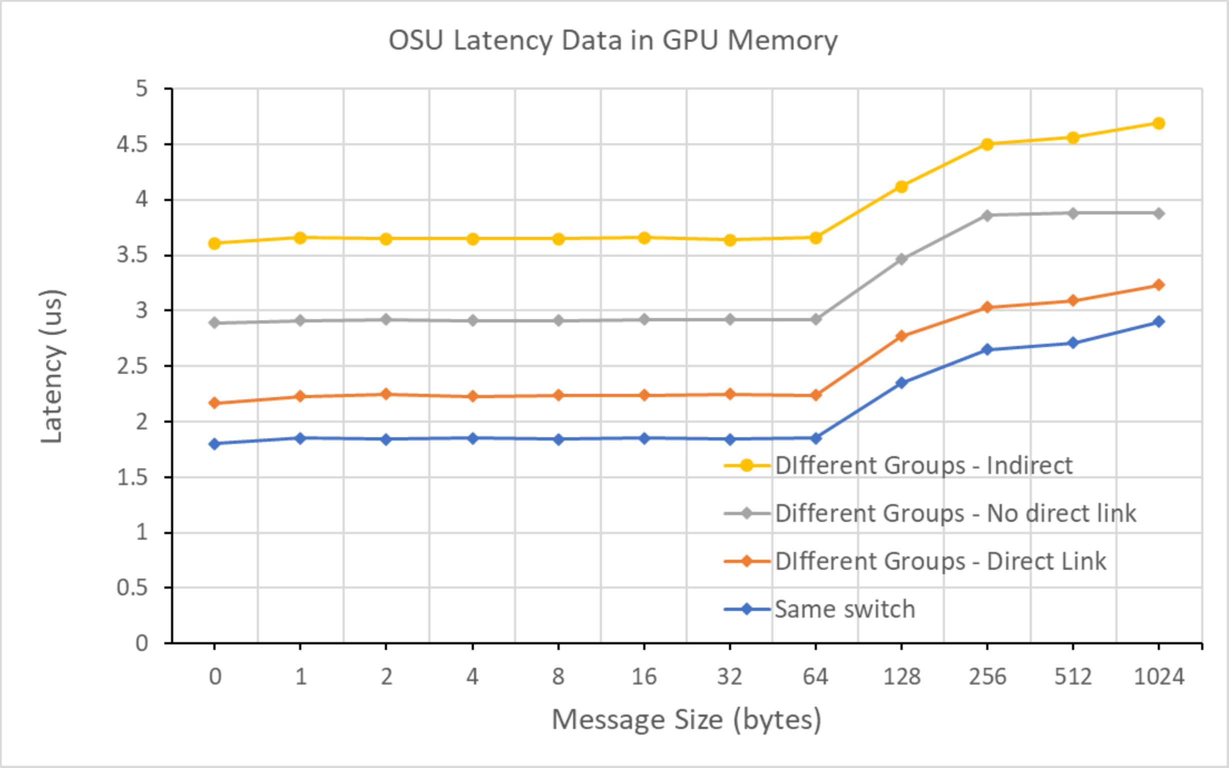
Another view of a Dragonfly Topology

- A group of endpoints connected to switches that are connected all-to-all
- A set of groups that are connected all-to-all

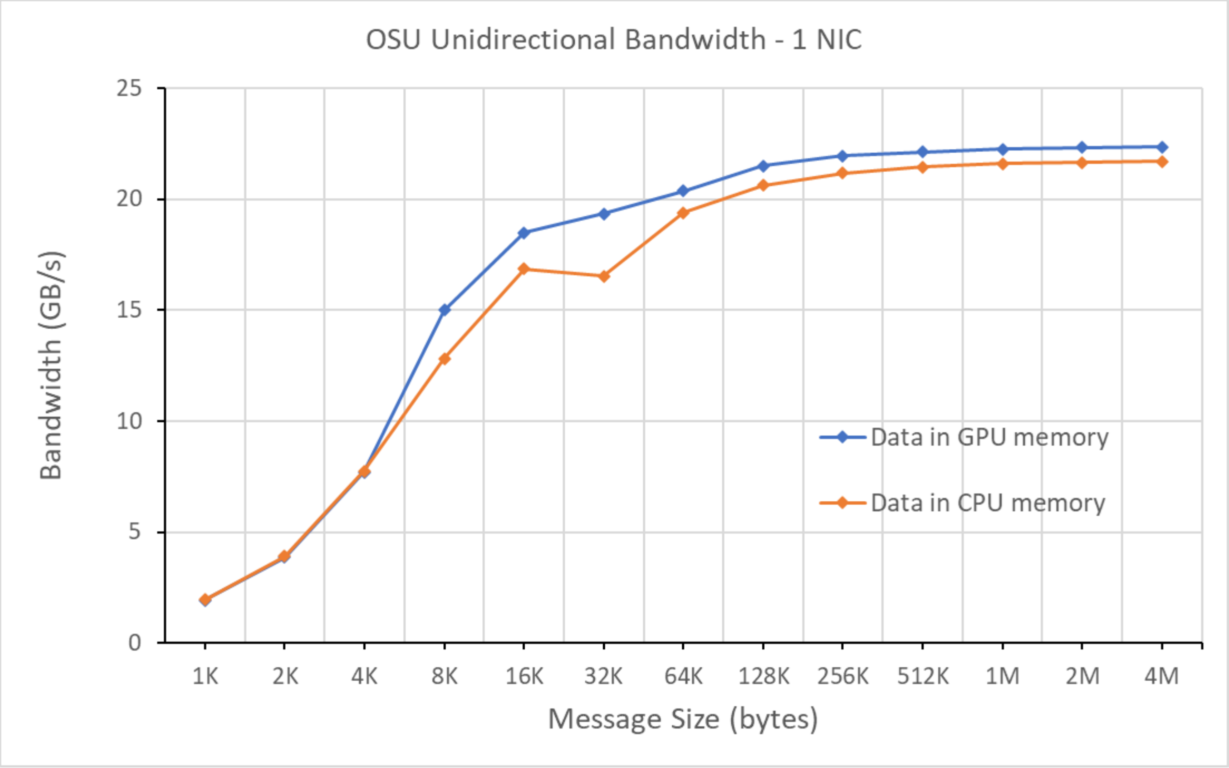


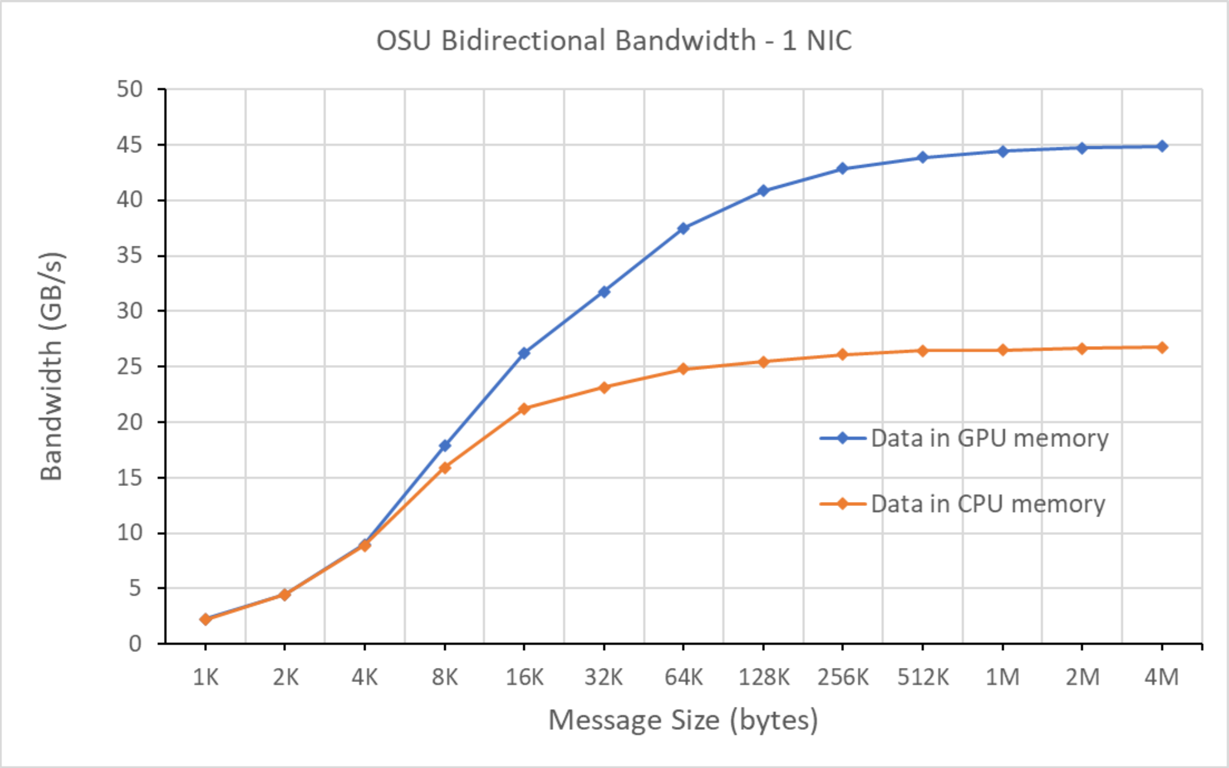
Similar Latency with CPU or GPU memory





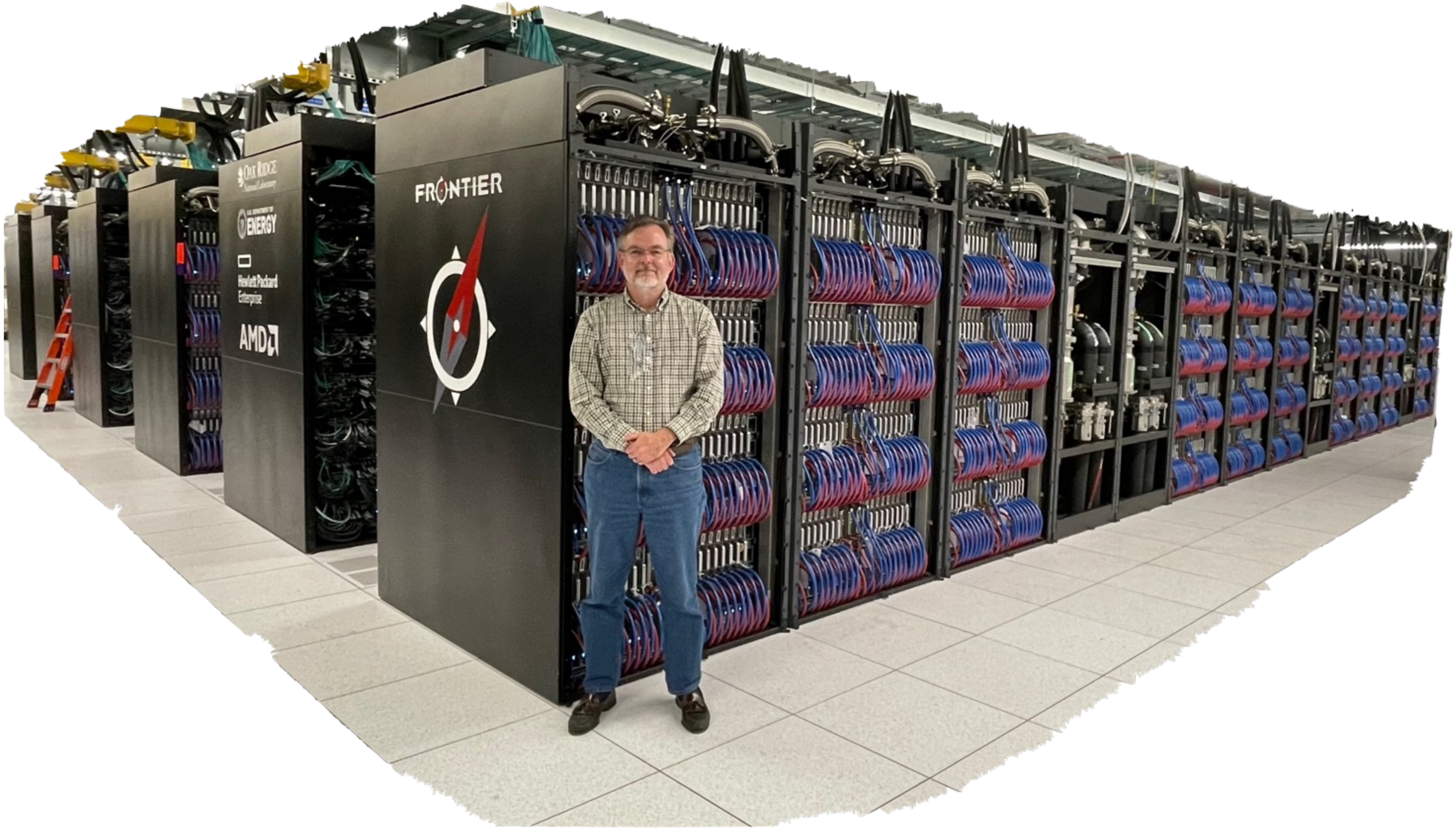
Better GPU Bandwidth





COPYRIGHT HPE 2022

Questions?


FRONTIER
OAK RIDGE National Laboratory
U.S. DEPARTMENT OF ENERGY
Hewlett Packard Enterprise
AMD