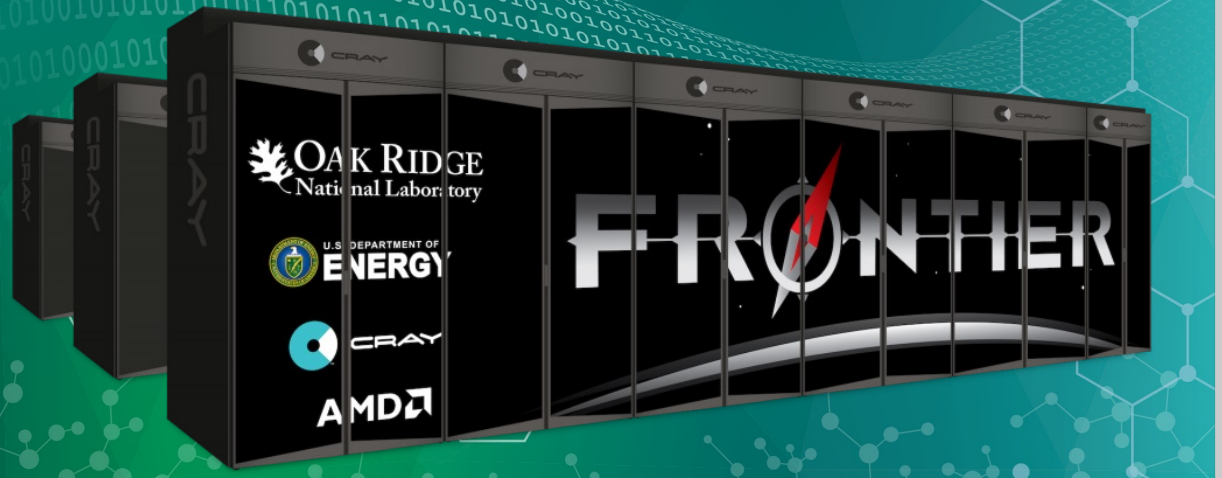
OAK RIDGE National Laboratory

OLCF's Frontier Supercomputer

ATPESC – July 31, 2023



Tom Papatheodore

HPC Engineer

System Acceptance & User Environment

Oak Ridge Leadership Computing Facility (OLCF)

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

U.S. DEPARTMENT OF ENERGY

DOE's Office of Science Computation User Facilities

NERSC

Argonne NATIONAL LABORATORY | Leadership Computing Facility

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Perlmutter: ~100 PF

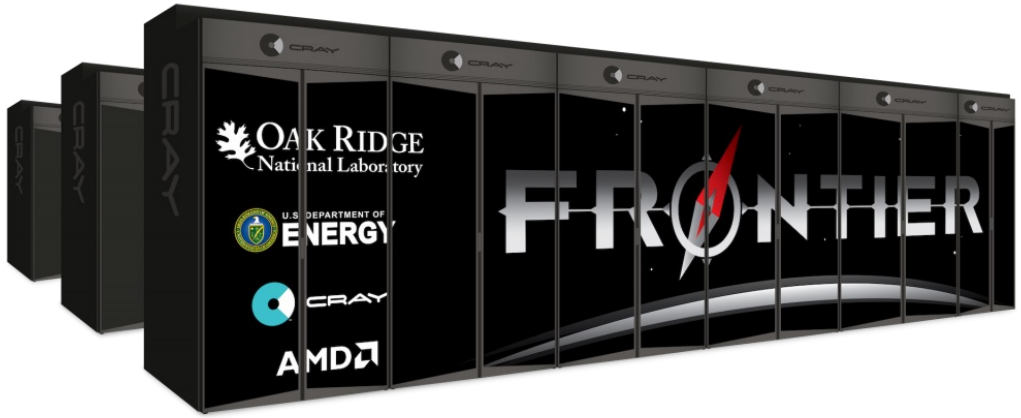
Aurora: >2 EF

Frontier: ~2 EF

- DOE is leader in open High-Performance Computing
- Provide the world's most powerful computational tools for open science
- Access is free to researchers who publish
- Boost US competitiveness
- Attract the best and brightest researchers

What is a Leadership Computing Facility (LCF)?

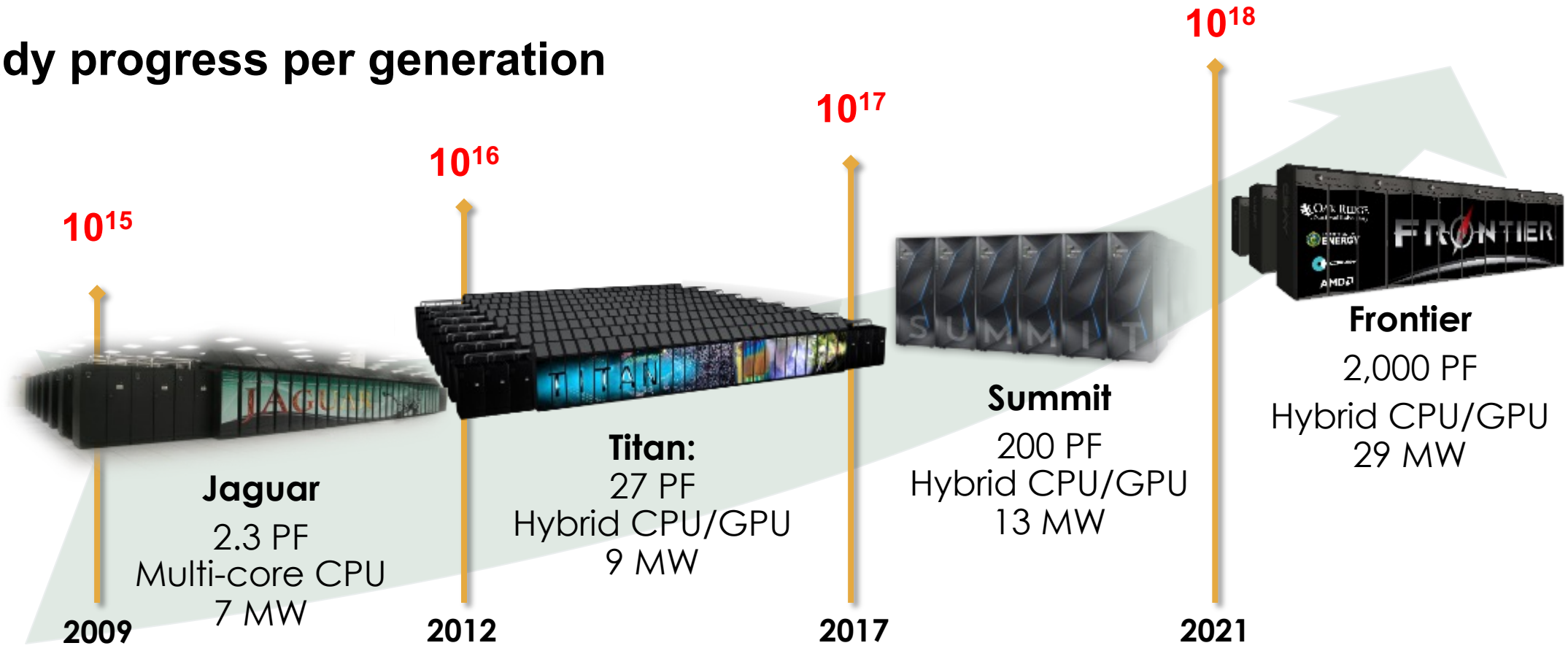
- Collaborative DOE Office of Science user-facility program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).



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

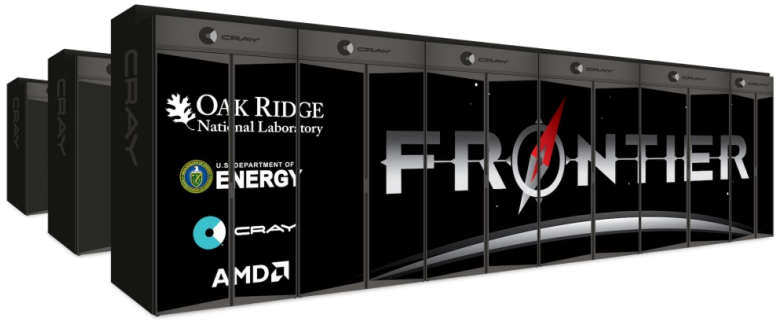


OLCF's Frontier Supercomputer


```
#1 Jun22 Top500 (1.102 EFlop/s)
#1 Nov22 Top500 (1.102 EFlop/s)
#1 Jun23 Top500 (1.194 EFlop/s)
```

Frontier Overview

Extraordinary Engineering



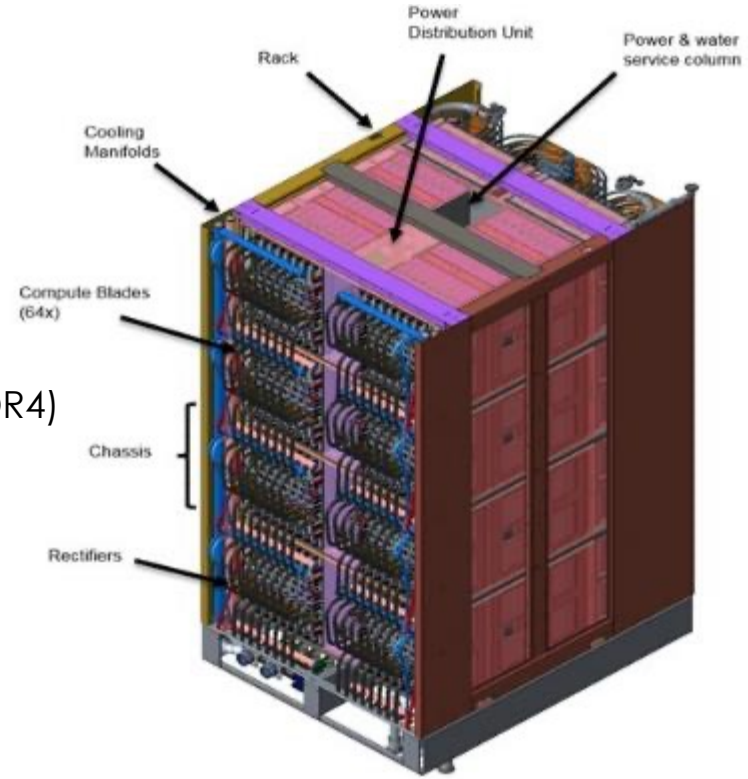
System

- 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

Built by HPE

Olympus rack

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



Powered by AMD

AMD node

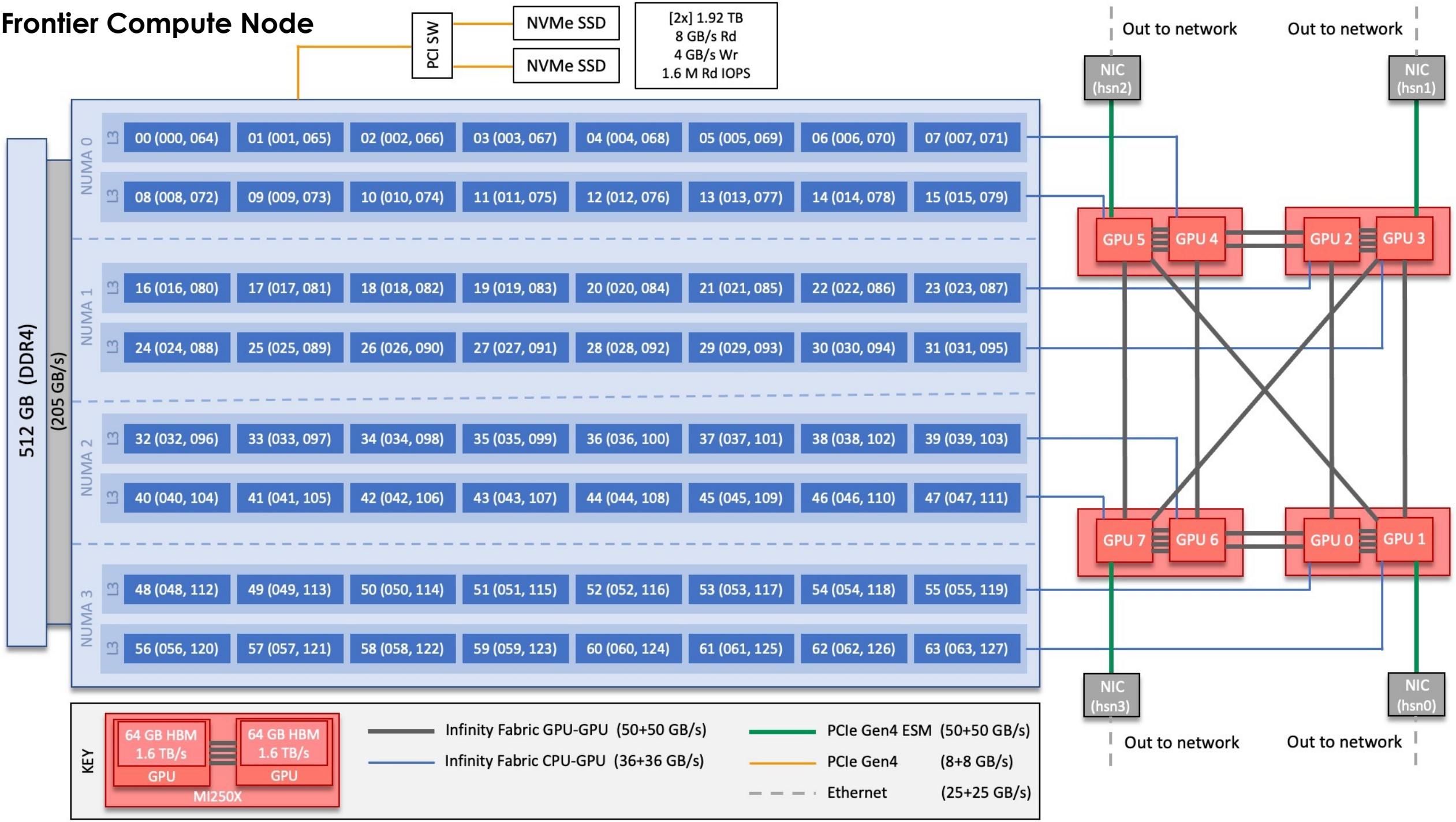
- 1 AMD "Optimized 3rd Gen EPYC" CPU
- 4 AMD MI250X GPUs
- 512 GiB DDR4 memory on CPU
- 512 GiB HBM2e total per node
  (128 GiB HBM per GPU)
- Coherent memory across the node
- 4 TB NVM
- GPUs & CPU fully connected with AMD Infinity Fabric
- 4 Cassini NICs, 100 GB/s network BW

Compute blade

- 2 AMD nodes

All water cooled, even DIMMS and NICs

Frontier Compute Node



OLCF Supercomputers: 2 Generations Later

- 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

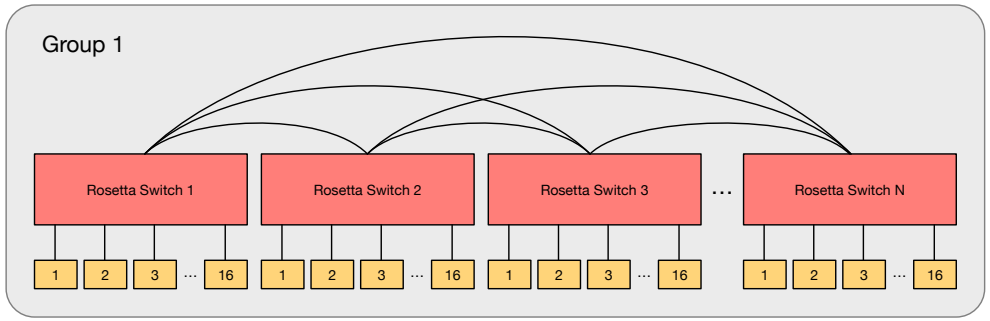
System	Titan (2012)	Summit (2017)	Frontier (2021)
Peak	27 PF	200 PF	2.0 EF
# nodes	18,688	4,608	9,408
Node	1 AMD Opteron CPU 1 NVIDIA Kepler GPU	2 IBM POWER9™ CPUs 6 NVIDIA Volta GPUs	1 AMD EPYC "Trento" CPU 4 AMD Instinct MI250X GPUs
Memory	0.6 PB DDR3 + 0.1 PB GDDR	2.4 PB DDR4 + 0.4 PB HBM + 7.4 PB NVM	4.6 PB DDR4 + 4.6 PB HBM2e + 36 PB NVM
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
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 8 GB/s 8 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 32:1 limited by PCIe	~~16:1~~ GPU:CPU 18:1 slightly limited by NVLink	64:1 GPU:CPU not limited by xGMI-2

What is Slingshot?

- HPC Ethernet Protocol
  - A superset of Ethernet
  - Negotiated between switch and NIC
    - Otherwise falls back to standard Ethernet
- Hardware
  - Rosetta switches
  - Cassini NICs
    - Accessed via OpenFabrics (aka libfabric)

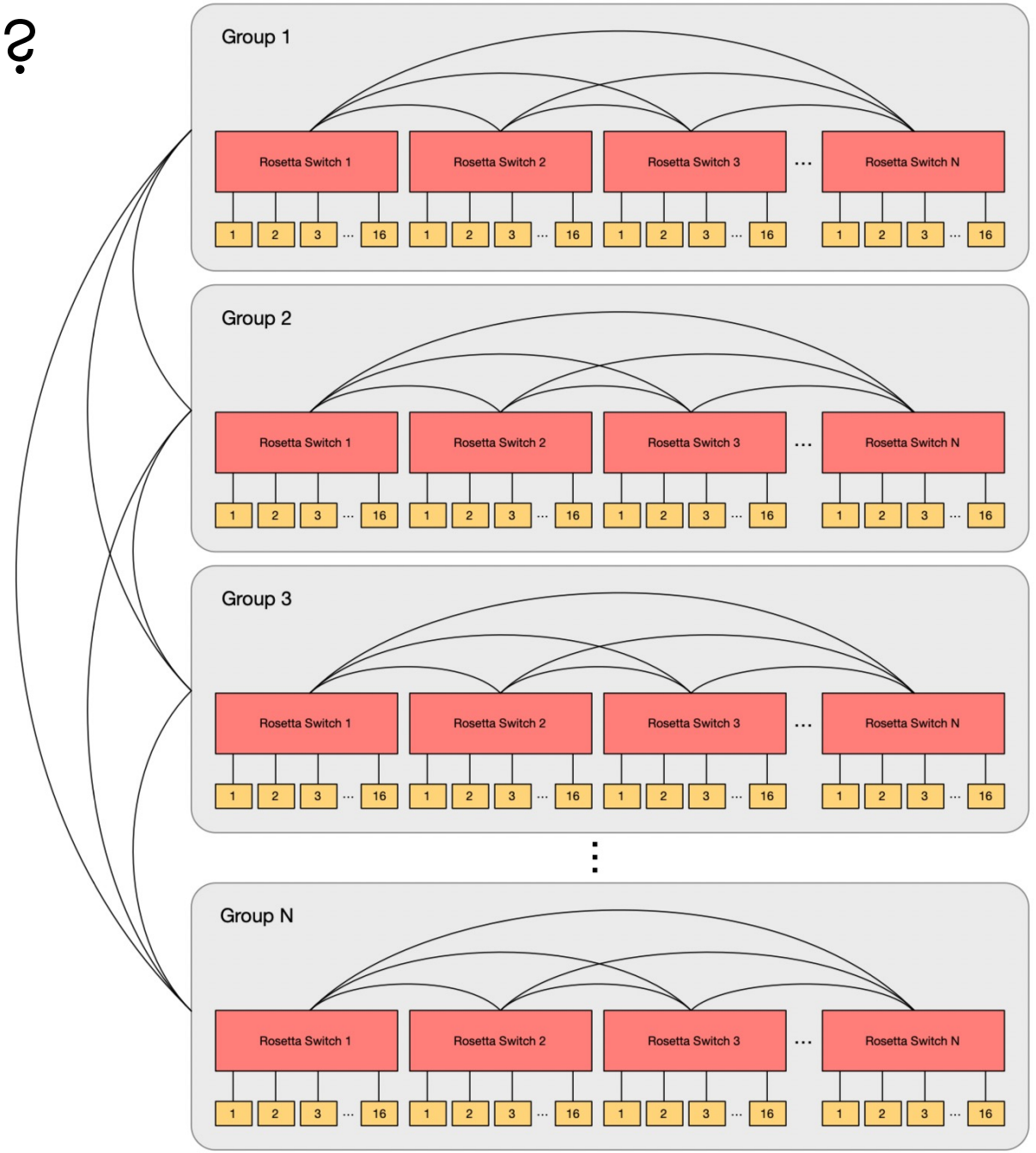
What is a Dragonfly group?

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



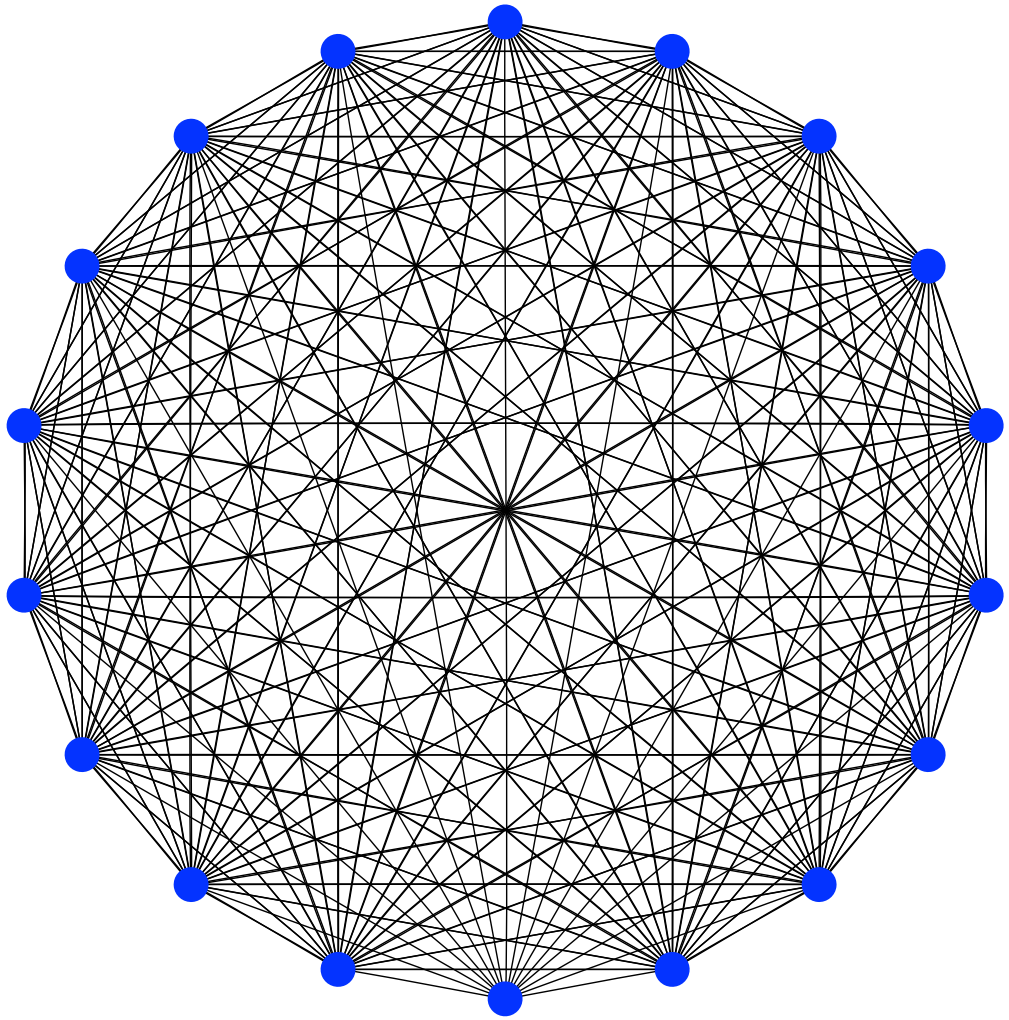
What is a Dragonfly topology?

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



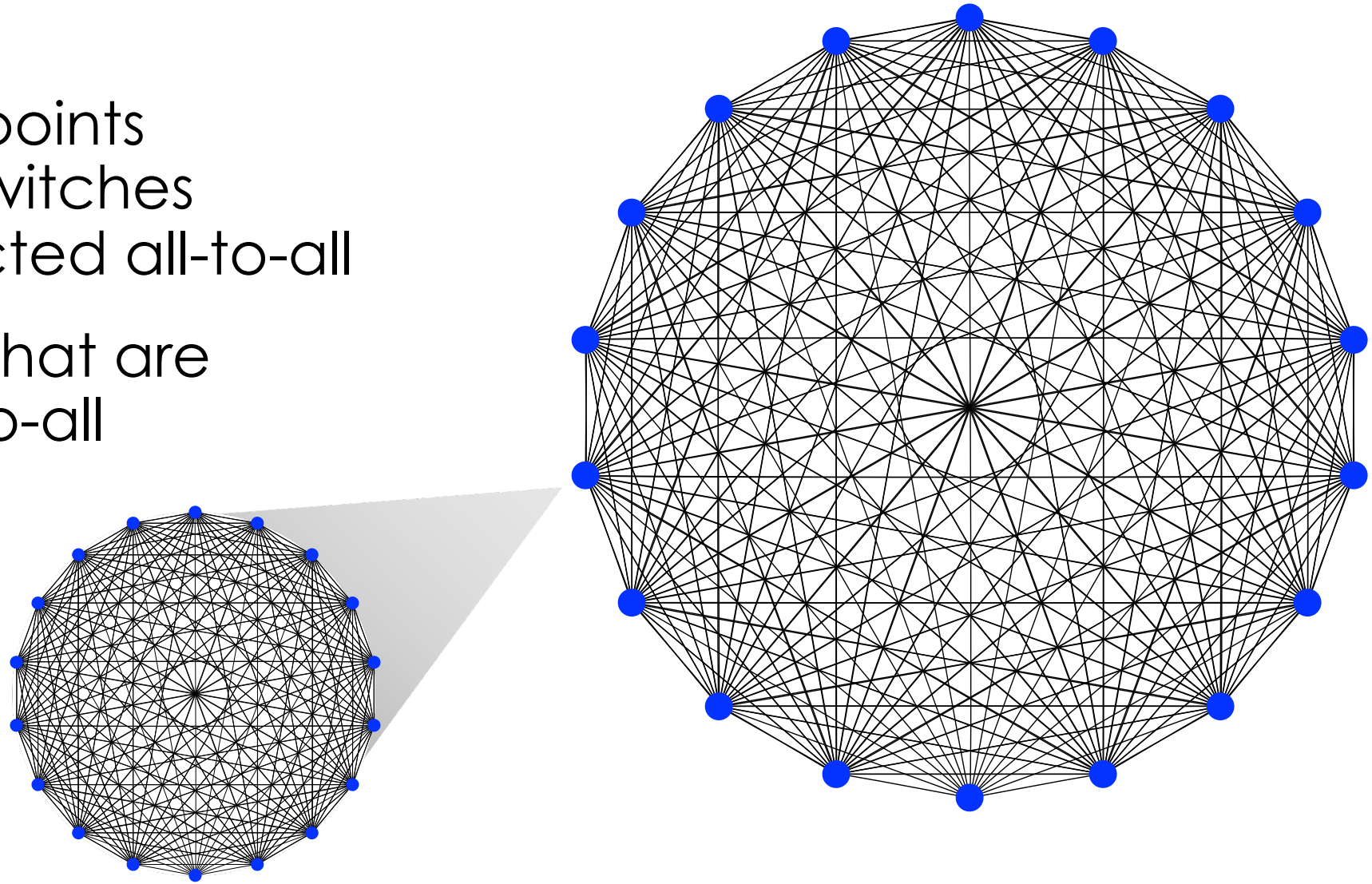
Another view of a Dragonfly Group

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



Another view of a Dragonfly Topology

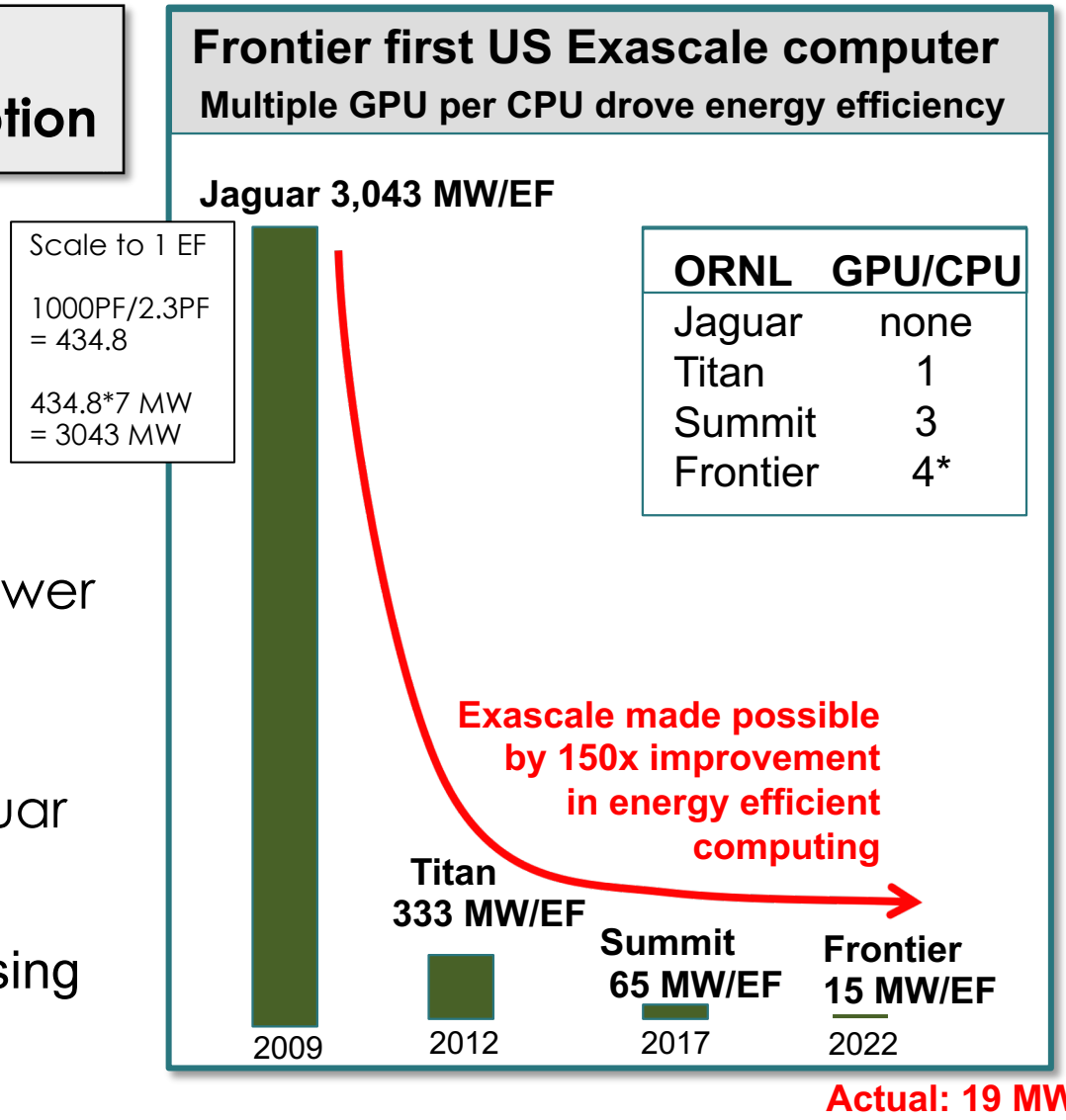
- A group of endpoints connected to switches that are connected all-to-all
- A set of groups that are connected all-to-all



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



OAK RIDGE National Laboratory

Questions?

papatheodore@ornl.gov

OAK RIDGE National Laboratory