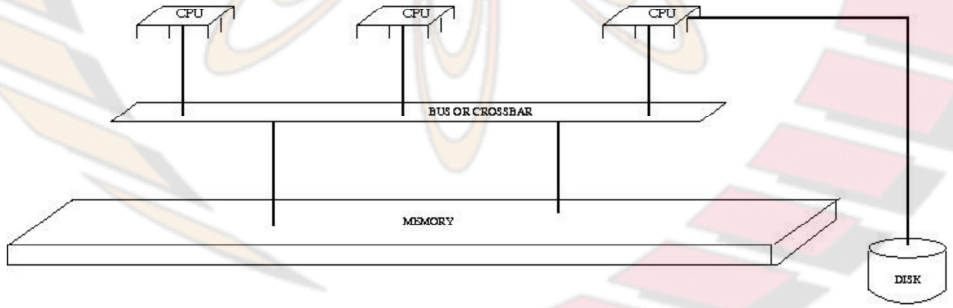
High Performance Scientific computing

S. Gopalakrishnan

Shared-Memory Processing

Each processor can access the entire data space

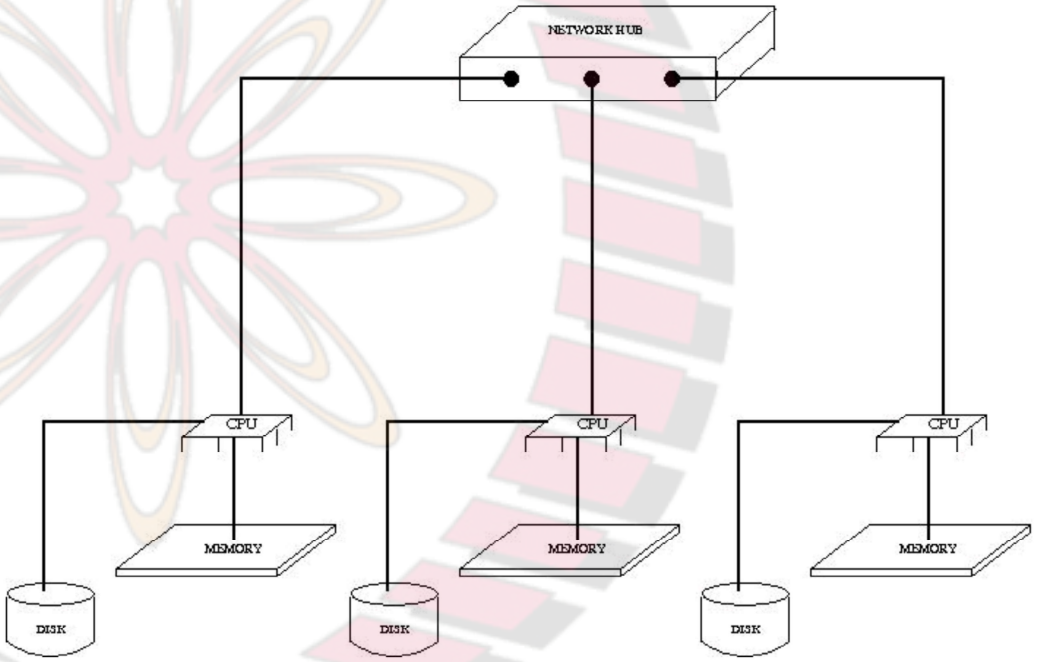
- Pro's
  - Easier to program
  - Amenable to automatic parallelism
  - Can be used to run large memory serial programs
- Con's
  - Expensive
  - Difficult to implement on the hardware level
  - Processor count limited by contention/coherency (currently around 512)
  - Watch out for "NU" part of "NUMA"



Distributed – Memory Machines

- Each node in the computer has a locally addressable memory space
- The computers are connected together via some high-speed network
  - Infiniband, Myrinet, Giganet, etc..

- Pros
  - Really large machines
  - Size limited only by gross physical considerations:
    - Room size
    - Cable lengths (10's of meters)
    - Power/cooling capacity
    - Money!
  - Cheaper to build and run
- Cons
  - Harder to program
    Data Locality



MPPs (Massively Parallel Processors)

Distributed memory at largest scale. Often shared memory at lower hierarchies.

- IBM BlueGene/L (LLNL)
  - 131,072 700 Mhz processors
  - 256 MB or RAM per processor
  - Balanced compute speed with interconnect

- Red Storm (Sandia National Labs)
  - 12,960 Dual Core 2.4 Ghz Opterons
  - 4 GB of RAM per processor
  - Proprietary SeaStar interconnect

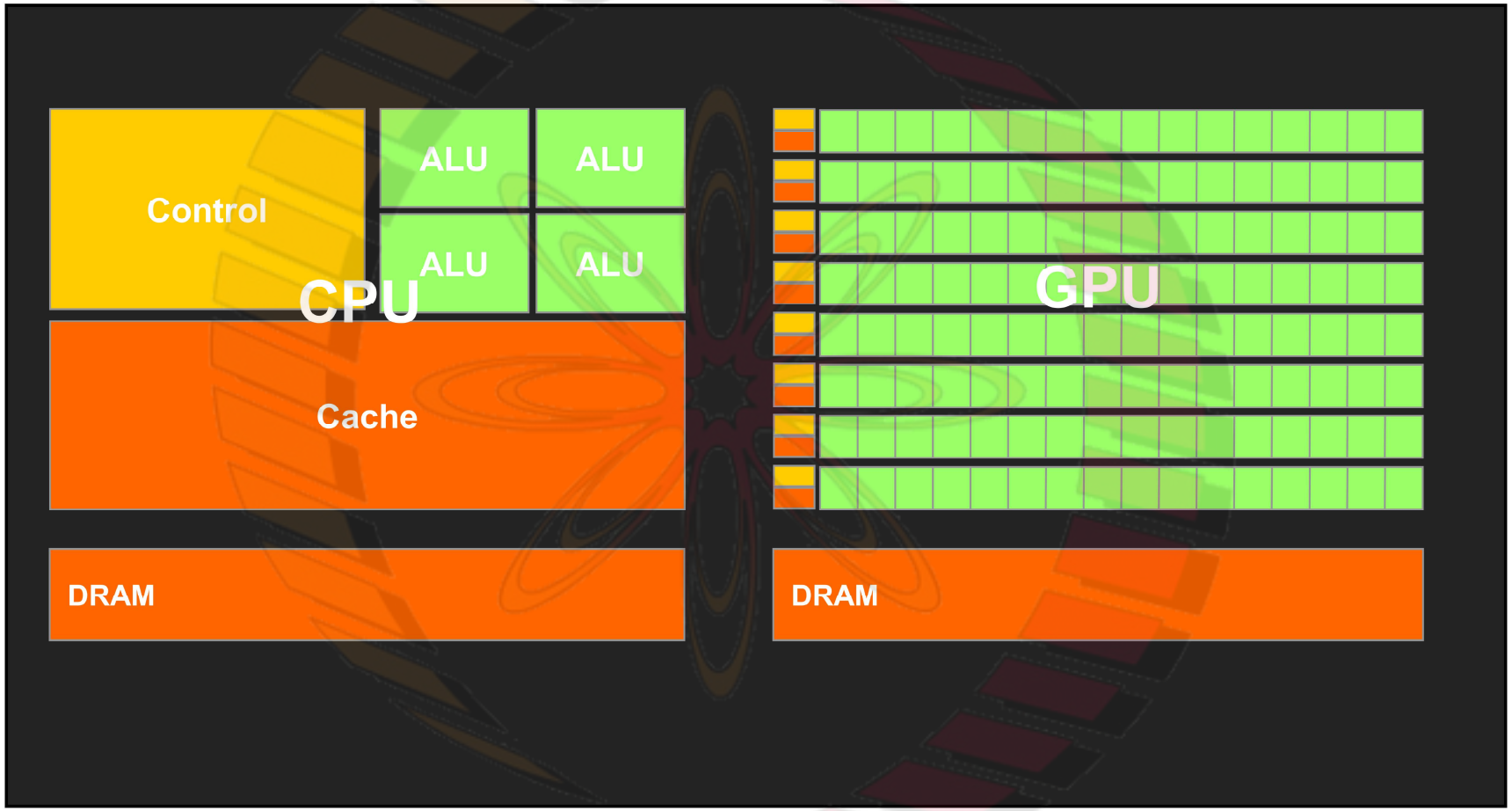
Frontier - World's fastest SC

Located at Oak Ridge National Labs , Tennessee, USA
Peak FLOPs - 1.7 ExaFlops (Peak)
Power Consumed - 23 MW
Cores - 8,699,904 - AMD EPYC 64C 2GHz

Spacetime2 - IIT Bombay

Located at Old CSE Building
Peak FLOPs - 1 PetaFlop (Peak)
Power Consumed - 0.2 MW
Memory - 43.75 TB
Cores - 9792 - Intel (Skylake + Broadwell), 64 P100 GPUS

Comparison of CPU vs GPU Architecture



Source: Prof. Wen-mei W. Hwu UIUC

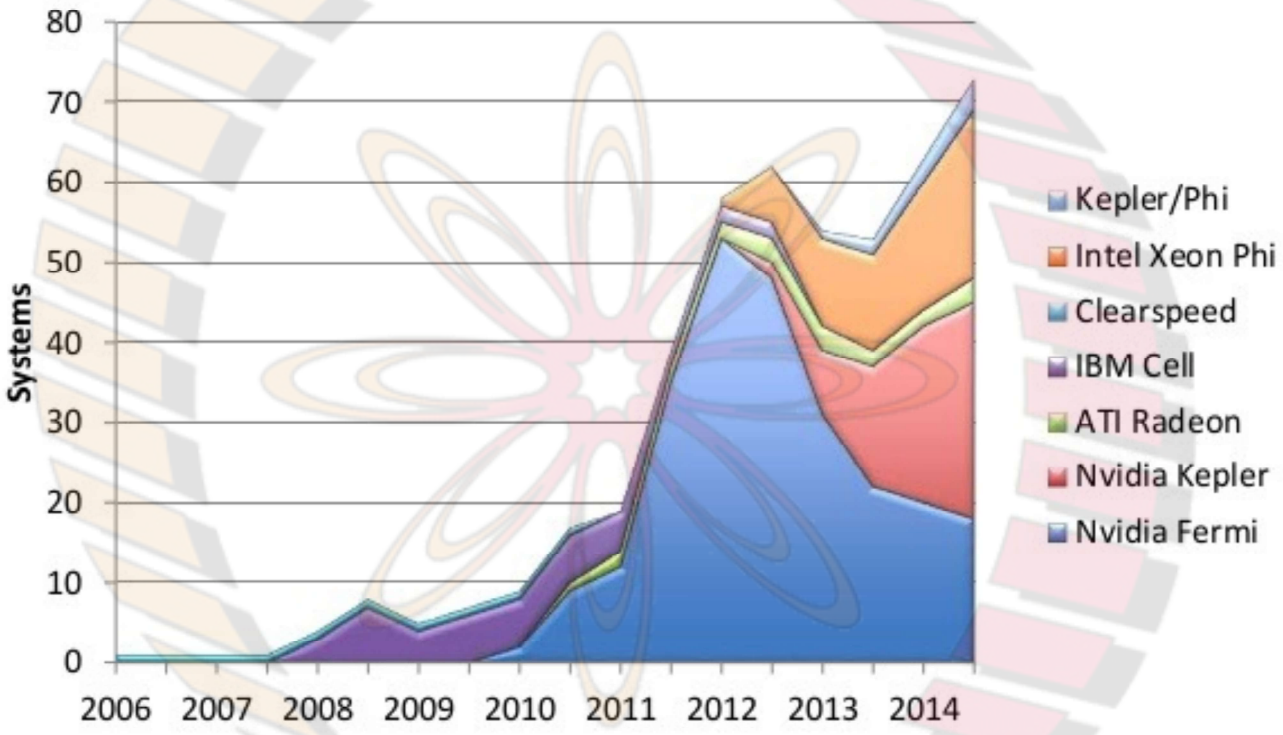
GPU CPU Analogy

GPU

CPU

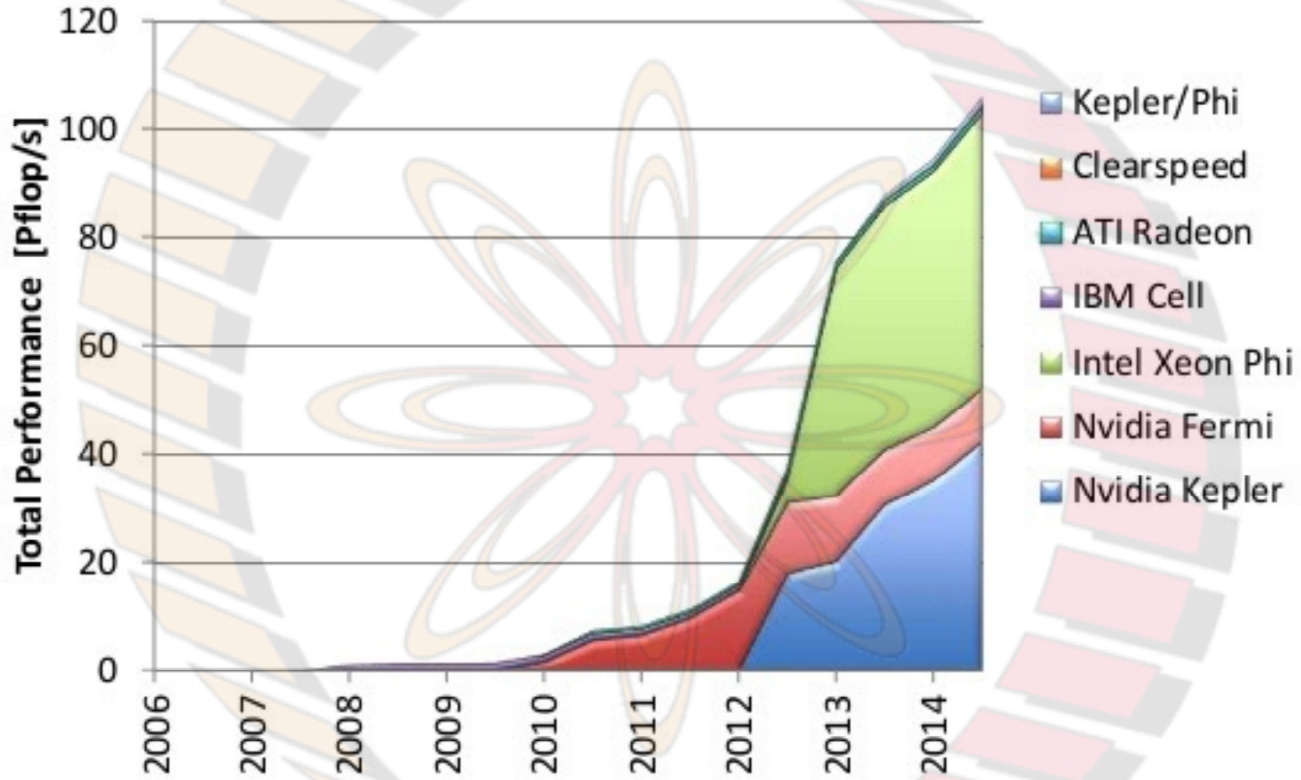
It is more effective to deliver Pizza's through light duty scooters rather than big truck. Similarly effective to use several lightweight GPU processors for parallel tasks.

ACCELERATORS



source: top500.org

PERFORMANCE OF ACCELERATORS



source: top500.org

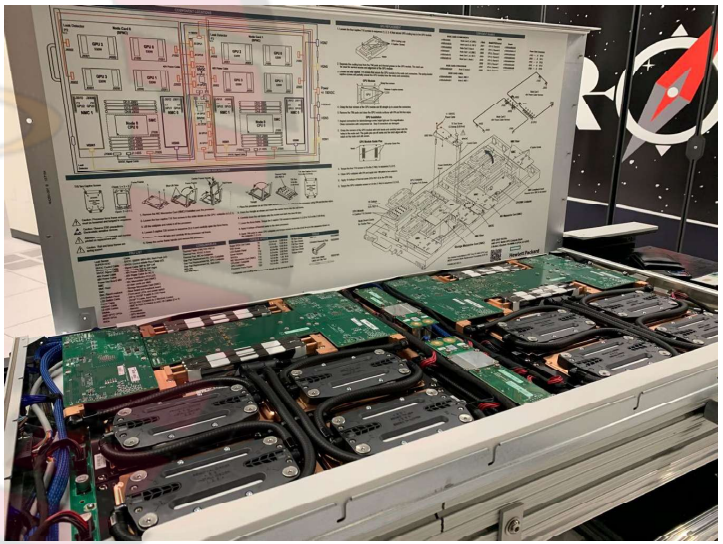
Frontier System Overview

- HPE Cray EX Supercomputer architecture
  - 9,408 compute nodes across 74 cabinets
  - 1.7 EF peak double-precision performance
    - 1.102 EF HPL performance (June 2022 debut)
    - 1.206 EF HPL performance(June 2024)
  - AMD 64-core Optimized 3rd Gen EPYC CPUs
  - AMD Instinct MI250X GPUs
  - HPE Slingshot interconnect
  - Cray and AMD ROCm prog. environments
  - 679 PB Lustre filesystem, "Orion"
  - NFS storage (/ccs/home, etc.)

https://www.flickr.com/photos/olcf/53567220071/

HPE Cray EX 235A Node Design

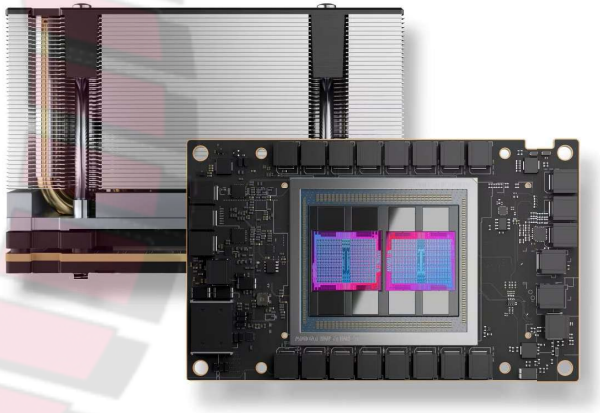
- 2x nodes per blade
  - Direct liquid cooled
- Each node has:
  - 1x AMD Optimized 3rd Gen EPYC CPU
  - 4x AMD Instinct MI250X GPUs
  - 512 GB DDR4 on CPU
  - 512 GB HBM2e per node
  - 2x 1.92 TB NVMe, "Burst Buffer"
  - Full CPU & GPU connectivity with AMD Infinity Fabric
  - 4x HPE Slingshot 200 GbE NICs



https://www.hpcwire.com/wp-content/uploads/2022/06/ORNL-Frontier-blade-display-closer-June2022_4000x-scaled.jpg

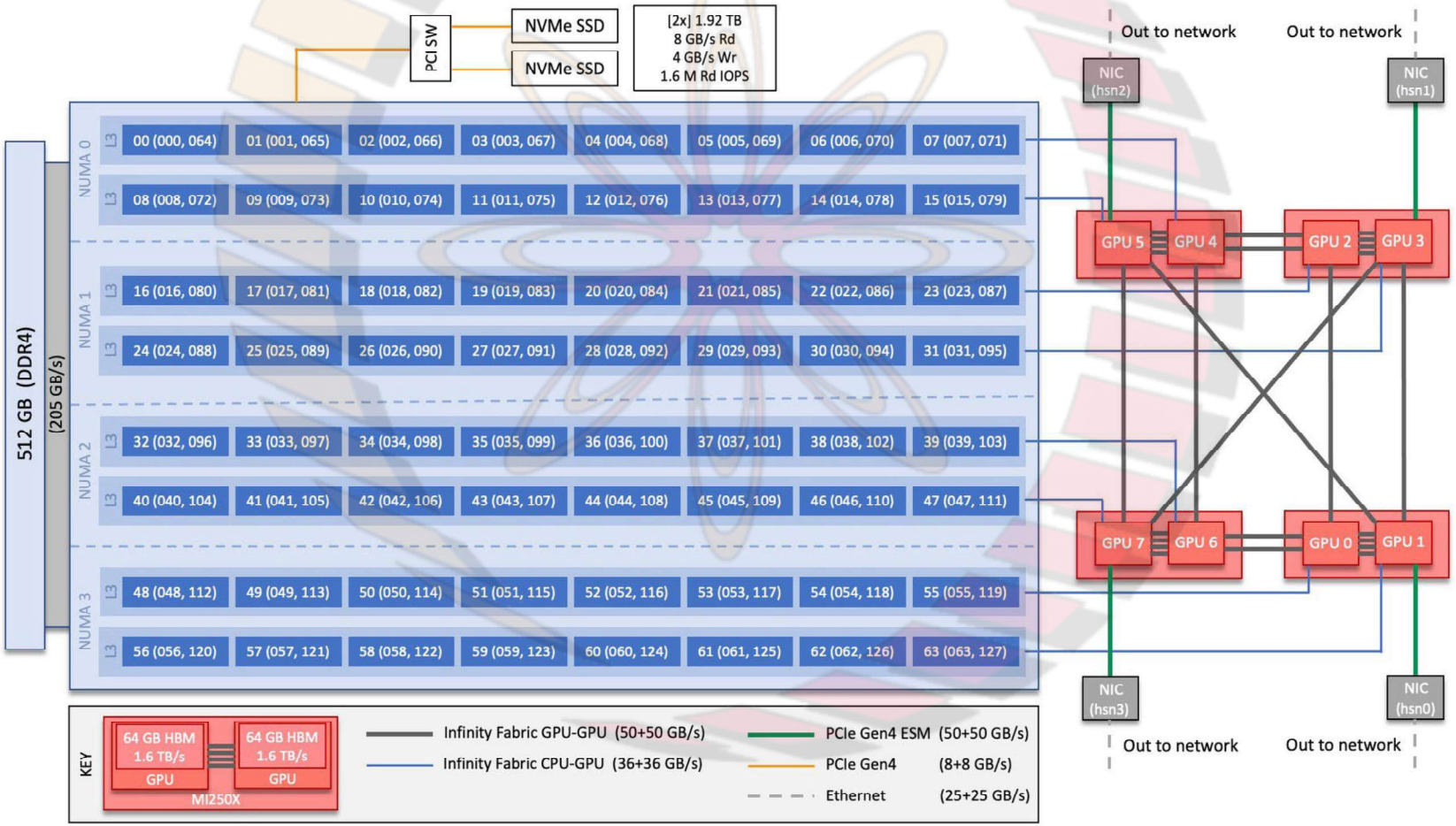
AMD Instinct MI250X GPU

- 2 Graphics Compute Dies (GCDs) per GPU
  - Shown by OS as 2 GPUs
- Effectively 8 GPUs per node, each with:
  - 110 Compute Units
  - 26.5 TFLOPs double-precision peak
  - 64 GB of HBM
  - 1.6 TB/s Memory Bandwidth
- Each associated with a CPU L3 cache region
- 1 NIC connected to each MI250X



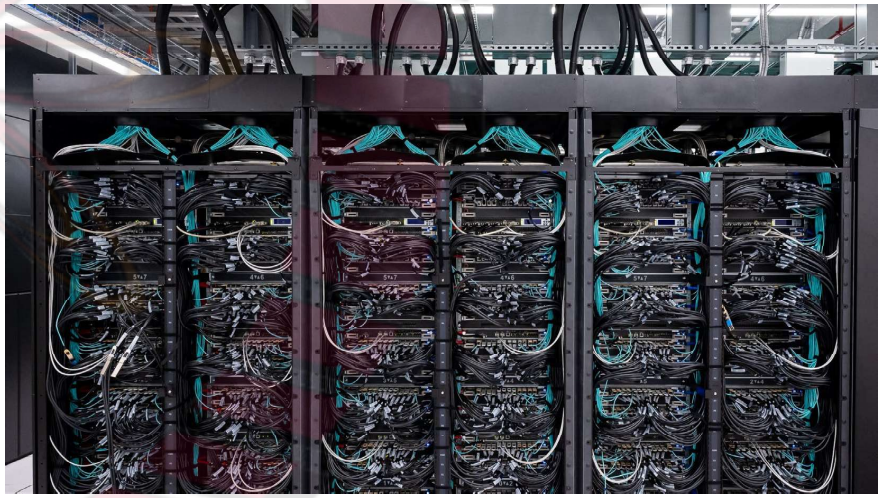
https://www.amd.com/content/dam/amd/en/images/products/data-centers/2325906-amd-instinct-mi250x-product.jpg

Frontier Node Diagram



HPE Slingshot Interconnect

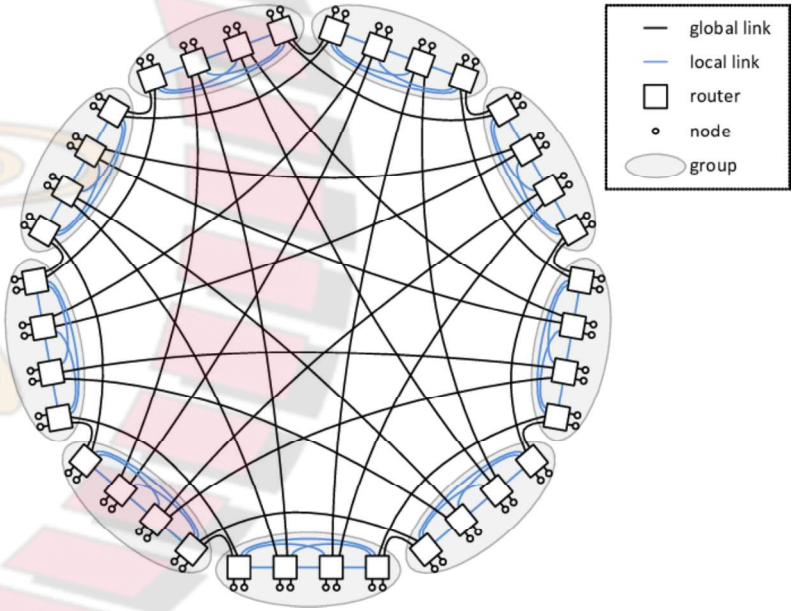
- High-speed, low latency network architecture
- HPE Slingshot switches (64 ports)
  - 25 GB/s bi-directional BW per port
- HPE Slingshot NICs
  - 25 GB/s bi-directional BW per link
- Slingshot is a superset of Ethernet with optimized HPC functionality
- Frontier uses dragonfly topology



https://www.ornl.gov/sites/default/files/2022-05/Side%20view%20Frontier%20cabinets.jpg

Frontier Network Topology

- Dragonfly groups
  - A group of endpoints connected to switches that are connected all-to-all
- Dragonfly topology
  - A set of groups connected all-to-all
    - Each group has >=1 link to every other group
- Frontier has 74 compute groups
  - 128 nodes per compute group
  - 32 switches per computer group
  - 4 NICs per node



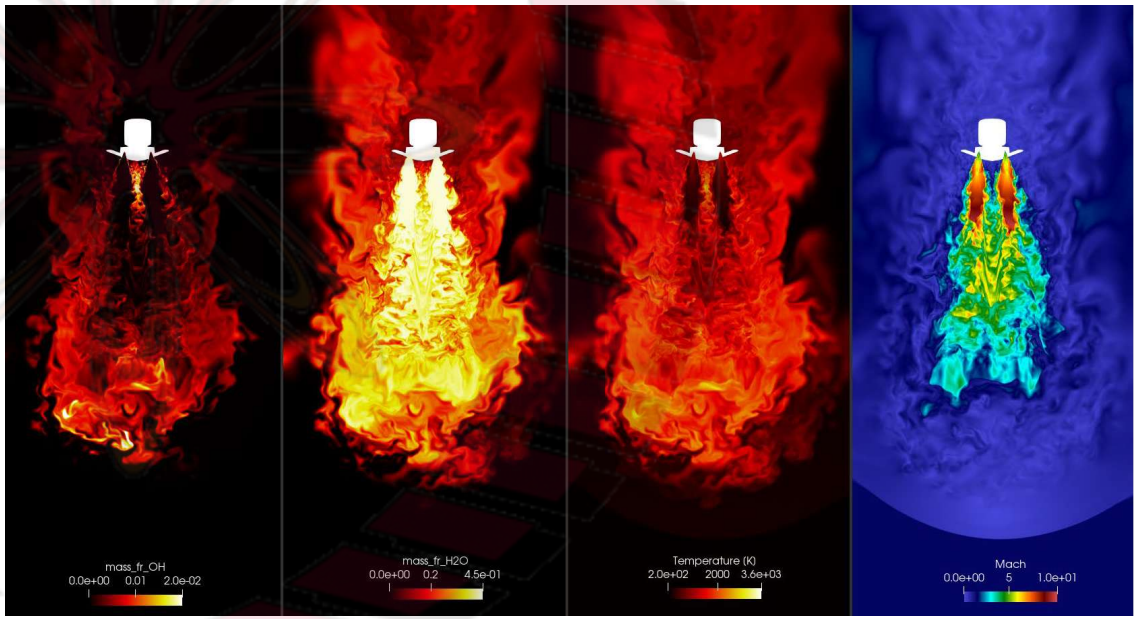
https://www.researchgate.net/profile/Enrique-Vallejo-2/publication/261313973/figure/fig2/AS:667782257573894@1536223105142/Sample-Dragonfly-topology-with-h2-p2-a4-36-routers-and-72-compute-nodes.png

Frontier Programming Environment

- Compilers
  - Cray CCE
    - C/C++ LLVM-based
    - Cray Fortran
  - AMD ROCm
    - C/C++ LLVM-based
  - GCC
  - oneAPI DPC++
    - LLVM-based
    - user-managed
- Programming Models & Abstraction Layers
  - OpenMP
  - HIP
  - Kokkos
  - RAJA
  - SYCL
    - via user-managed DPC++
  - OpenACC
    - C/C++ via user-managed clacc
  - OpenCL
  - UPC++

Example Frontier Use Case

- NASA is exploring ways to safely land a vehicle bringing humans to Mars
- Unable to flight-test in Martian environment
- Frontier enabled first-of-kind test flights
  - New levels of resolution, physical modeling, and temporal duration

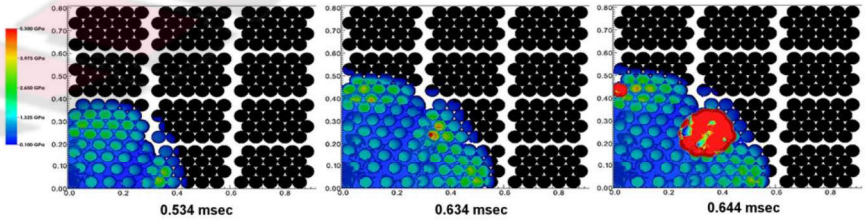


https://www.olcf.ornl.gov/wp-content/uploads/static_fine-1500x811.png

Example Leadership-Class Use Case

- Semi hauling 35,000 pounds of mining explosives crashed in Utah
- Caught fire and caused dramatic explosion leaving a 30'x70' crater
- Debris launched up to 1/4 mile
- Leadership-class systems (e.g., Titan) used to recreate explosion
- Uintah simulations helped identify safer ways to pack explosives

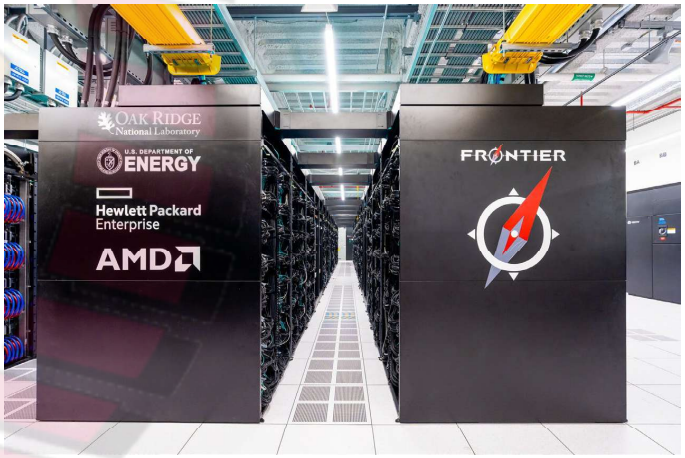
https://www.summitdaily.com/news/trailer-full-of-explosives-blows-hole-in-utah/



https://www.jics.tennessee.edu/files/images/accidental-explosion1.jpg

Fun Facts

- 1 Exaflop => 10^{18} Calculations per Second
  - Frontier can do in 1 second what'd take over 4 years if everyone on Earth did 1 calculation/s
- Theoretical peak of 2 Exaflop
  - Compute similar to 194,544 PS5s
- 74 cabinets weighing 8,000 pounds each
  - 1 cabinet has 10% more performance than Titan
    - Using 309 kW compared to Titan's 7 MW
- 700 PB of storage
  - 25 Mt. Everests of DVDs



https://www.flickr.com/photos/olcf/52117839159/