OAK RIDGE National Laboratory

Frontier: Breaking the Exascale Barrier with the Fastest Supercomputer Ever

Bronson Messer
Director of Science
Leadership Computing Facility
Oak Ridge National Laboratory

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

U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

Oak Ridge Leadership Computing Facility (OLCF)

Mission: Deploy and operate the computational and data resources required to tackle global challenges

- Providing the resources to investigate otherwise inaccessible systems at every scale: from galaxy formation to supernovae to earth systems to automobiles to nanomaterials

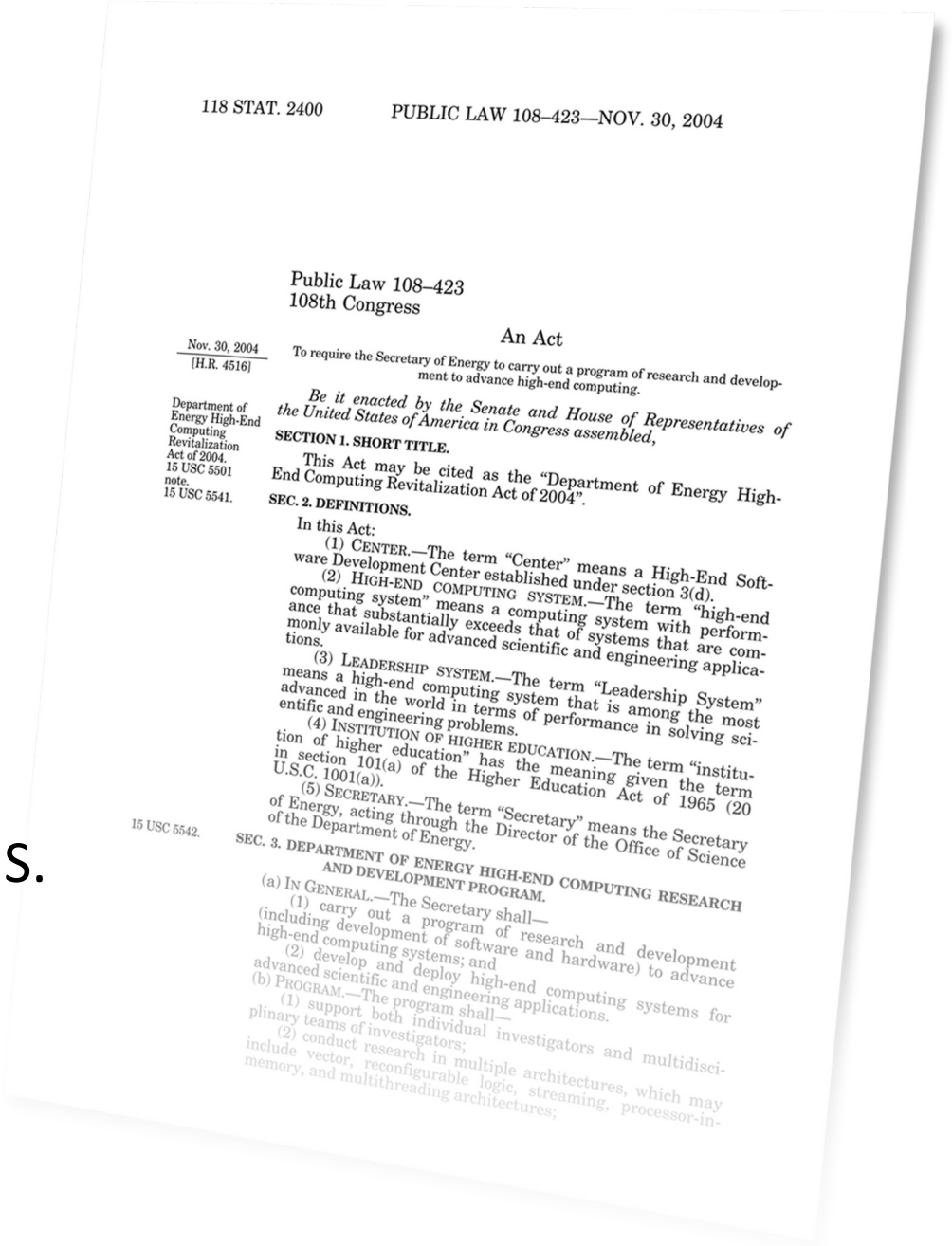
- With our partners, deliver transforming discoveries in materials, biology, climate, energy technologies, and basic science

Leadership Computing Facilities

Department of Energy High-End Computing Revitalization Act of 2004 (Public Law 108-423):

The Secretary of Energy, acting through the Office of Science, shall

- Establish and operate Leadership Systems Facilities
- Provide access [to Leadership Systems Facilities] on a competitive, merit-reviewed basis to researchers in U.S. industry, institutions of higher education, national laboratories and other Federal agencies.



118 STAT. 2400 PUBLIC LAW 108–423—NOV. 30, 2004

Public Law 108–423
108th Congress

An Act

Nov. 30, 2004
[H.R. 4516]

To require the Secretary of Energy to carry out a program of research and development to advance high-end computing.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

Department of Energy High-End Computing Revitalization Act of 2004.
15 USC 5501 note.

SECTION 1. SHORT TITLE.

This Act may be cited as the "Department of Energy High-End Computing Revitalization Act of 2004".

15 USC 5541.

SEC. 2. DEFINITIONS.

In this Act:

(1) CENTER.—The term "Center" means a High-End Software Development Center established under section 3(d).

(2) HIGH-END COMPUTING SYSTEM.—The term "high-end computing system" means a computing system with performance that substantially exceeds that of systems that are commonly available for advanced scientific and engineering applications.

(3) LEADERSHIP SYSTEM.—The term "Leadership System" means a high-end computing system that is among the most advanced in the world in terms of performance in solving scientific and engineering problems.

(4) INSTITUTION OF HIGHER EDUCATION.—The term "institution of higher education" has the meaning given the term in section 101(a) of the Higher Education Act of 1965 (20 U.S.C. 1001(a)).

(5) SECRETARY.—The term "Secretary" means the Secretary of Energy, acting through the Director of the Office of Science of the Department of Energy.

15 USC 5542.

SEC. 3. DEPARTMENT OF ENERGY HIGH-END COMPUTING RESEARCH AND DEVELOPMENT PROGRAM.

(a) IN GENERAL.—The Secretary shall—

(1) carry out a program of research and development (including development of software and hardware) to advance high-end computing systems; and

(2) develop and deploy high-end computing systems for advanced scientific and engineering applications.

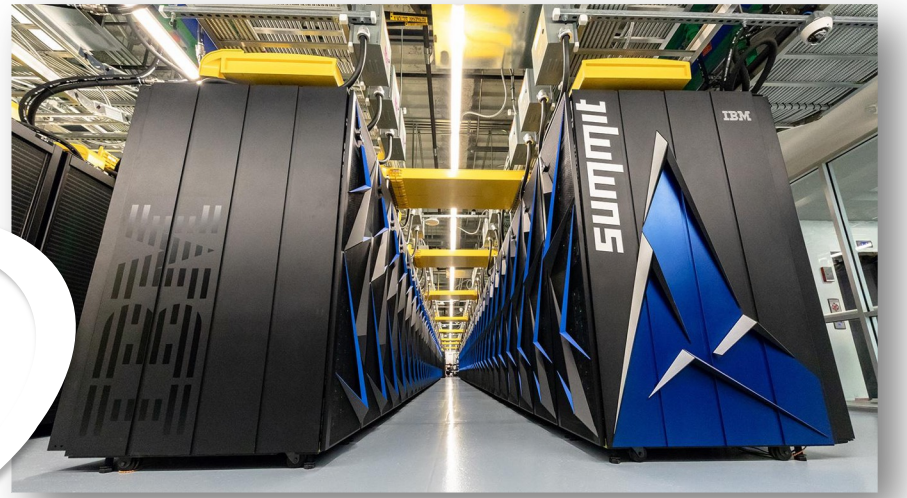
(b) PROGRAM.—The program shall—

(1) support both individual investigators and multidisciplinary teams of investigators;

(2) conduct research in multiple architectures, which may include vector, reconfigurable logic, streaming, processor-in-memory, and multithreading architectures;

What is the Leadership Computing Facility (LCF)?

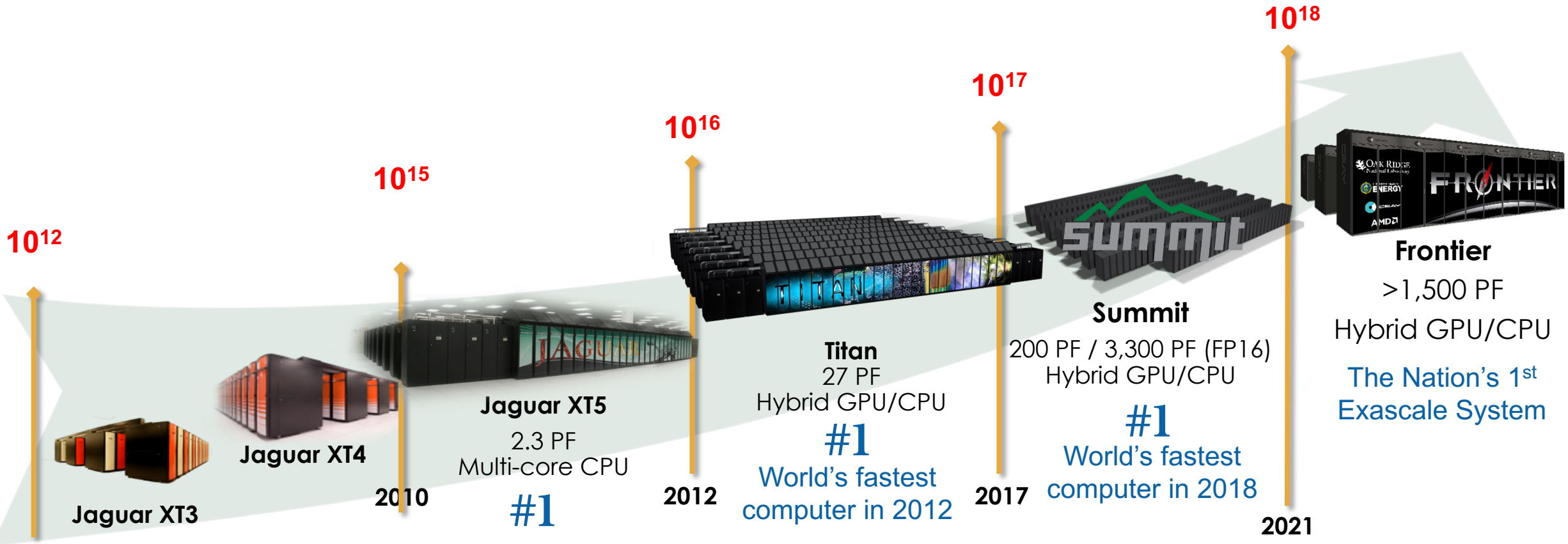
- Collaborative DOE Office of Science user-facility program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2 centers/2 architectures to address diverse and growing computational needs of the scientific community
- Highly competitive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



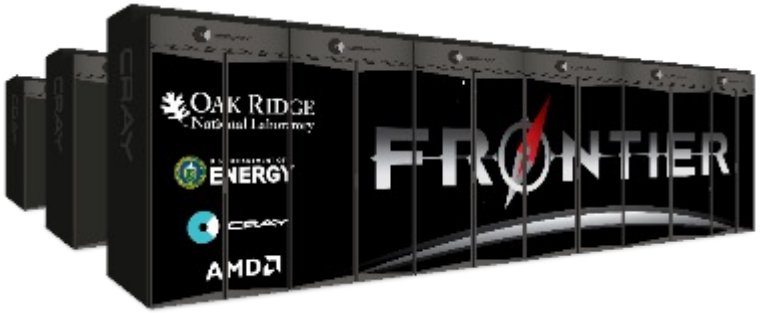
OAK RIDGE National Laboratory

Argonne NATIONAL LABORATORY

ORNL has had a Top 10 supercomputer in every year since the Leadership Computing Facility was founded in 2005. Jaguar, Titan, and Summit are the only DOE/SC systems to be ranked #1 on the TOP500 list of fastest computers.



Frontier Overview



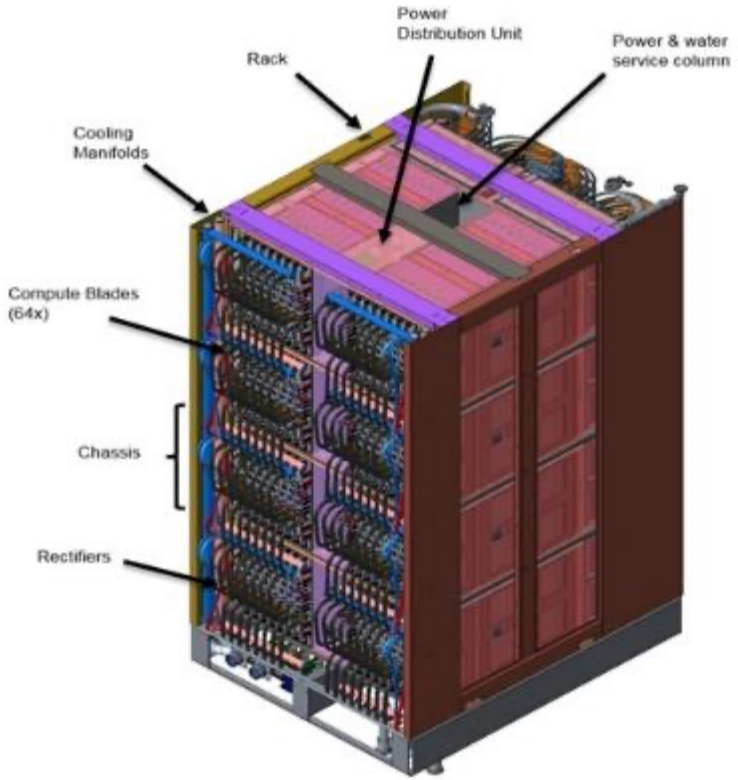
System

- 2 EF Peak DP FLOPS
- 74 compute racks
- 29 MW Power Consumption
- 9,408 nodes
- 9.2 PB memory
  (4.6 PB HBM, 4.6 PB DDR4)
- Cray Slingshot network with dragonfly topology
- 37 PB Node Local Storage
- 716 PB Center-wide storage
- 4000 ft^2 foot print

Built by HPE

Olympus rack

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



Powered by AMD

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

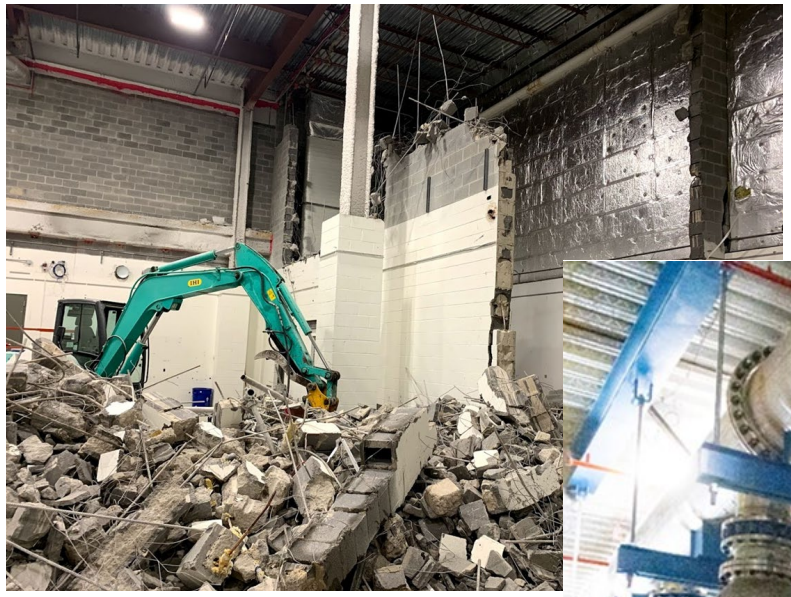
Compute blade

- 2 AMD nodes

All water cooled, even DIMMS and NICs

Power, space, and cooling – (one of) the hard part(s)

- 30 offices, 8 laboratories, and a 20,000 s.f. data center were repurposed



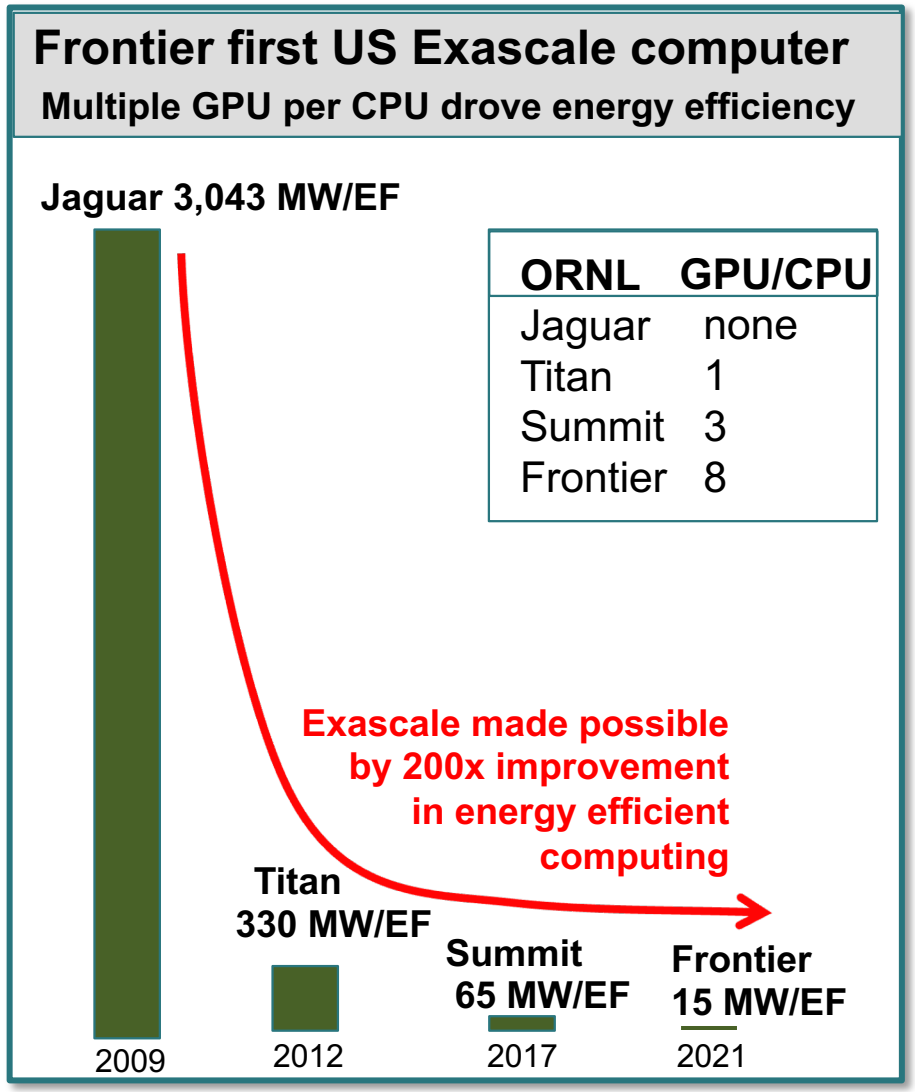
40 MW of power

A new data center (recall the 8,000 lb cabinets…)

Energy-efficient computing – Frontier achieves 14.5 MW per EF

Since 2009 the biggest concern 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.
- **ASCR [Fast, Design, Path] Forward vendor investments** in energy efficiency (2012-2020) further reduced the power consumption of computing chips (CPUs and GPUs)..
- **200x 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



During Frontier build -- the chip shortage hit in earnest!

When HPE began ordering parts, suppliers said the lead time on orders was increasing an additional 6-12 months.

60 Million parts needed for Frontier

685 Different part numbers used in Frontier

167 Frontier part numbers affected by the chip shortage
(more than 2 million parts from dozens of suppliers worldwide)

12 Part numbers blocked building the first compute cabinet

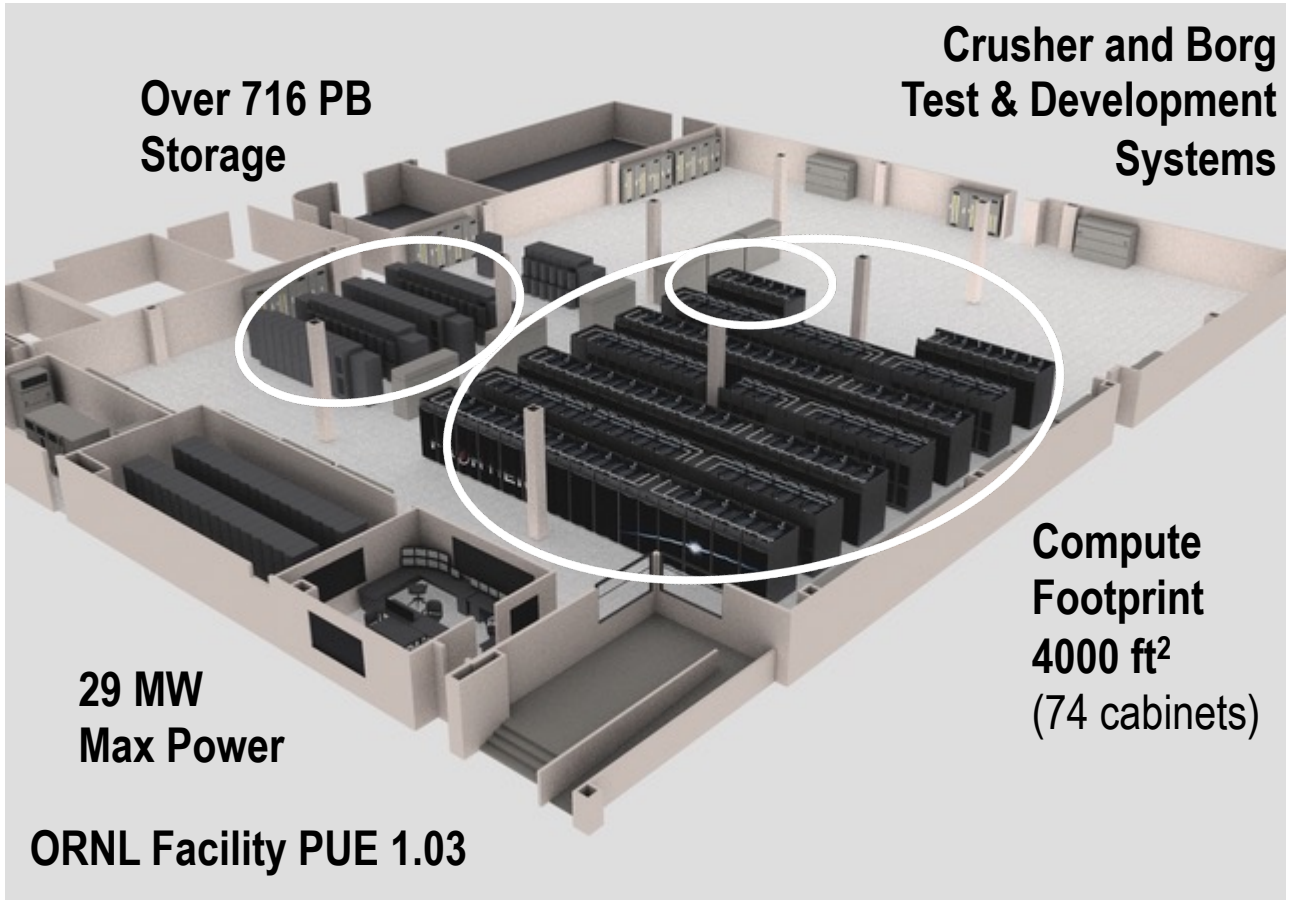
15 Part numbers shortage for AMD building all the MI200 cards for Frontier

It wasn't exotic parts like CPUs or GPUs, rather parts needed by everyone – in cars, TVs, electronics, such as voltage regulators, oscillators, power modules, etc.

Last Cabinet of Frontier Delivered to ORNL October 18, 2021

Thanks to Heroic Efforts of the HPE and AMD teams

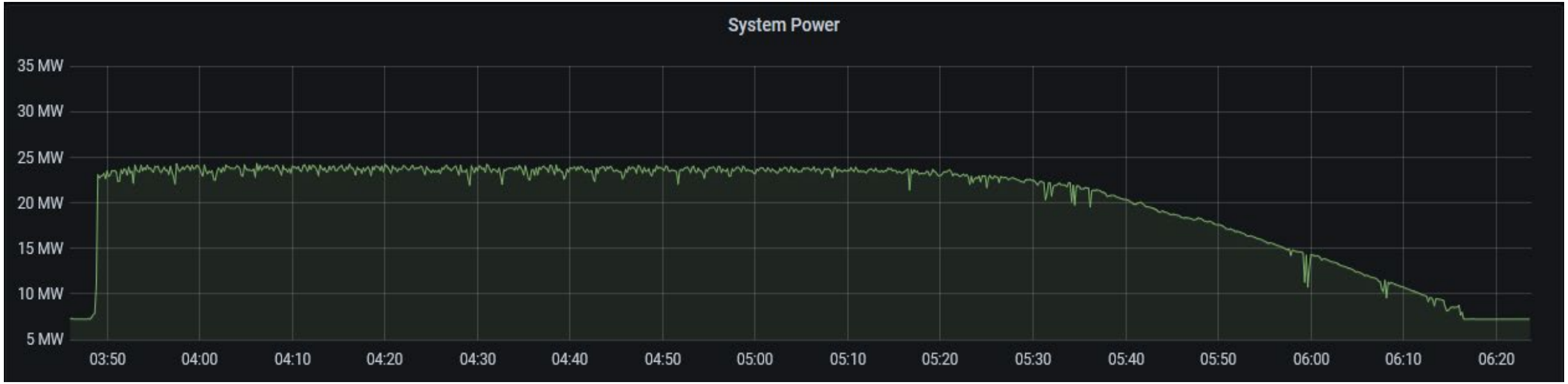
Last cabinet being rolled into place.
(Each cabinet weighs 8,000 lbs.)



After the cabinets arrived they had to be connected. There are 81,000 cables between all the Frontier nodes

Then system debug and tuning began

- We fell into a pattern of repairing hardware, updating software, and tuning the system by day
- And running benchmarks like HPL at night



- In May, as time was running out for the June Top500, we had a successful exascale HPL run:

9,248 nodes of Frontier achieved 1.1 EF
#1 TOP500 list
#2 Green500 achieving over 52 Gflop/W

OAK RIDGE NATIONAL LABORATORY'S FRONTIER SUPERCOMPUTER

- **74 HPE Cray EX cabinets**
- **9,408 AMD EPYC CPUs, 37,632 AMD GPUs**
- **700 petabytes of storage capacity, peak write speeds of 5 terabytes per second using Cray Clusterstor Storage System**
- **90 miles of HPE Slingshot networking cables**

TOP500

#1

1.1 exaflops of performance on the May 2022 Top500.

GREEN500

#1, #2

62.04 gigaflops/watt power efficiency on a single cabinet.
52.23 gigaflops/watt power efficiency on the full system.

HPL-AI

#1

6.88 exaflops on the HPL-AI benchmark.

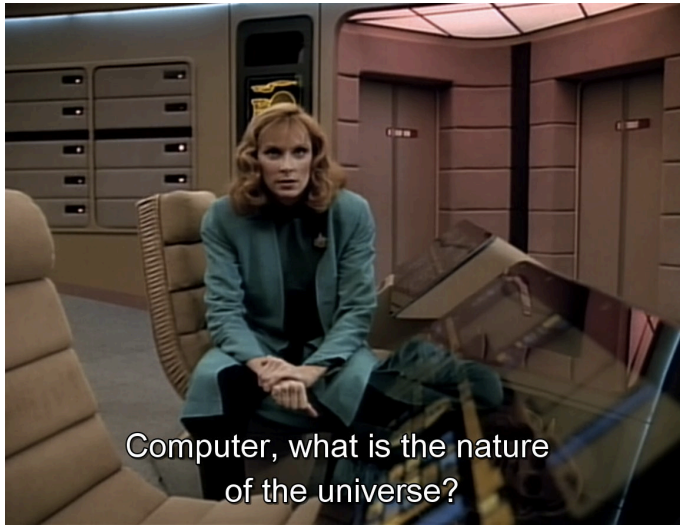
Sources: May 30, 2022 Top500 release

Frontier multi-tier storage system

Multi-tier I/O Subsystem	Capacity	Read	Write
	37 PB Node Local Storage (Two 2TB SSD NVM per node)	65.9 TB/s	62.1 TB/s
		11 Billion IOPS	
	11 PB Performance tier	9.4 TB/s	9.4 TB/s
	695 PB Capacity tier	5.2 TB/s	4.4 TB/s
	10 PB Metadata	2M Transactions per sec	

Crusher (Frontier Test and Development System)

- 2 cabinets, the first with 128 compute nodes and the second with 64 compute nodes, for a total of 192 compute nodes. ~40PF (!!)
  - *Crusher is about as powerful as 1.5 Titans!*
- Each node
  - One 64-core AMD EPYC 7A53 CPU
  - 512 GB of DDR4 memory.
  - Four AMD MI250X, each with 2 Graphics Compute Dies (GCDs) for a total of 8 GCDs per node
  - Connected with 4 HPE Slingshot 200 Gbps NICs
- Kept in rough sync with Frontier SW stack



CAAR

The **Center for Accelerated Application Readiness (CAAR)** is the primary OLCF program to achieve and demonstrate application readiness

- ***Build on the experience from the successful CAAR programs for OLCF-3 (Titan) and OLCF-4 (Summit)***
- ***CAAR project resources***
  - Dedicated collaboration with OLCF staff
  - Support and consultation from other project personnel, particularly from the Programming Environment and Tools area, and the vendor Center of Excellence
  - OLCF Postdoctoral fellows (both during application readiness and early science)
  - Allocations to available compute resources (Summit, early access systems)

The Center for Accelerated Application Readiness (CAAR)

- Built on the successful programs for OLCF-3 (Titan) & OLCF-4 (Summit)
- CAAR has been working with 8 applications since mid 2019 as part of OLCF-5
- Also supporting work on applications through ECP
- These applications have access to early hardware and software through the Vendor Center of Excellence

CAAR Applications		ECP Applications	
Astrophysics	CHOLLA	Astrophysics	ExaStar
Molecular Dynamics	NAMD	Astrophysics	ExaSky
		HEP	LatticeQCD
Materials Science	LSMS	Chemistry	NWCHEMeX
Biology/Health	CoMet	Chemistry	GAMESS
Fluid Dynamics	GESTS	Combustion	PELE
		Energy	ExaSMR
Nuclear Physics	NUCCOR	Energy	WDMApp
		Climate	E3SM
Plasma Physics	PIConGPU	Additive Manufacturing	ExaAM
Subsurface Flow	LBPM	Biology	ExaBiome
		Electric Grids	ExaSGD

Characteristics of CAAR Projects

Application	Programming languages	Scientific libraries used	I/O	Algorithms	Initial parallelization
Cholla	C++	None.	HDF5	Finite volume hydrodynamics	MPI, CUDA
NAMD	C++	FFTW (node-level)	VMD (custom)	MD, PME	CHARM++, CUDA
LSMS	F90/C++	BLAS, LAPACK, FFTW	HDF5	Dense Linear Solvers, Coupled ODE, Poisson Eq., Monte Carlo	MPI+CUDA
CoMet	C++	cuBLAS, MAGMA	None	2-way and 3-way Proportional Similarity Method and Custom Correlation Coefficient	MPI+OpenMP, CUDA
GESTS	F90	FFTW	HDF5	Fourier pseudo-spectral methods	MPI+OpenMP 4.5
NUCCOR	F90 + F2008; C	BLAS, LAPACK	HDF5	CCSD + CCSDT, Hartree-Fock, Sparse and dense linear algebra (eigensolvers)	MPI+OpenMP, CUDA
PIConGPU	C++	Alpaka, SOLLVE	ADIOS	PIC	MPI+OpenMP, CUDA/HIP/TBB thru Alpaka
LBPM	C++	Zlib	SILO, HDF5	Lattice Boltzmann methods	MPI, CUDA

ECP Application Portfolio – Early Science runs on Frontier

Earth system

Climate Change

Subsurface use for **carbon capture**, petroleum extraction, waste disposal

Accurate regional impact assessments in **Earth system models**

Stress-resistant crop analysis and catalytic conversion of **biomass-derived alcohols**

Metagenomics for analysis of biogeochemical cycles, climate change, environmental remediation

Energy security

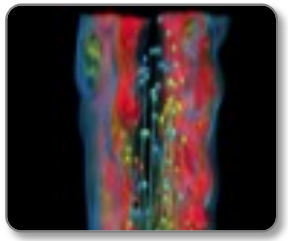
Reliable and efficient planning of the power grid

Turbine **wind plant** efficiency

Design and commercialization of **Small Modular Reactors**

Nuclear fission and fusion reactor **materials design**

High-efficiency, low-emission **combustion engine** and gas turbine design

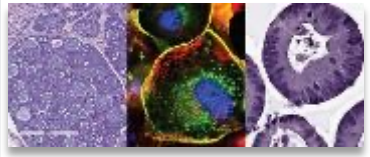


Health care

Accelerate and translate cancer research

(partnership with NIH)

Developing AI for Precision Drug Therapy in Fight Against Cancer



Scientific discovery

Cosmological probe of the standard model of particle physics

Validate fundamental laws of nature

Find, predict, and control materials and properties

Light source-enabled analysis of protein and molecular structure and design

Predict and control magnetically confined fusion plasmas

Demystify origin of chemical elements

Economic security

Additive manufacturing of qualifiable metal parts

Scale up of **clean fossil fuel** combustion

Biofuel catalyst design

Seismic hazard risk assessment

Large Scale Density Functional Theory at the Exascale with LSMS

Workflows and high performance computations to predict materials properties

Research Topics

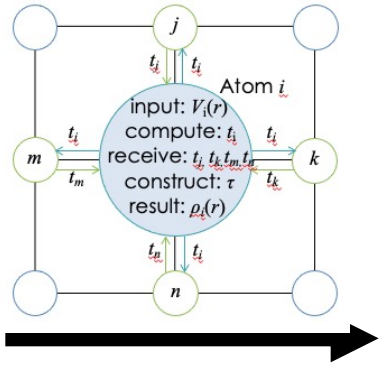
- Understanding the role of disorder and defects in materials for electronic and mechanical properties
- Complex magnetic order – topological magnetic structures (e. g. Skyrmions) and magnetism beyond ideal crystal

Recent Highlights

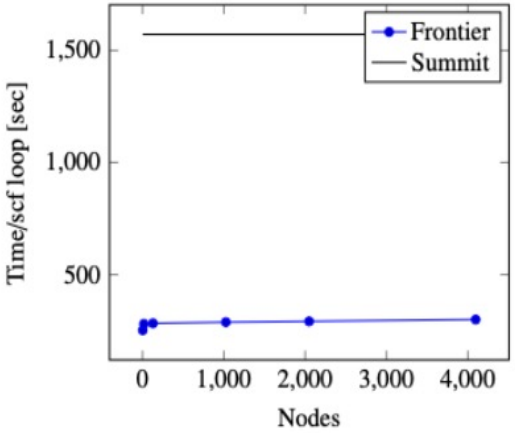
- Successful porting of the LSMS code (github.com/mstsuite/lsms) to Frontier for exascale materials simulations.
- Scaling of first principles calculations to O(100,000) up to O(1,000,000) atoms for the first time.
- Demonstrated scaling of LSMS on Frontier up to 1,048,576 atom FePt system on 8192 Frontier nodes.
- Speedup of LSMS from Summit to Frontier from combined hardware and software improvements is ~8x

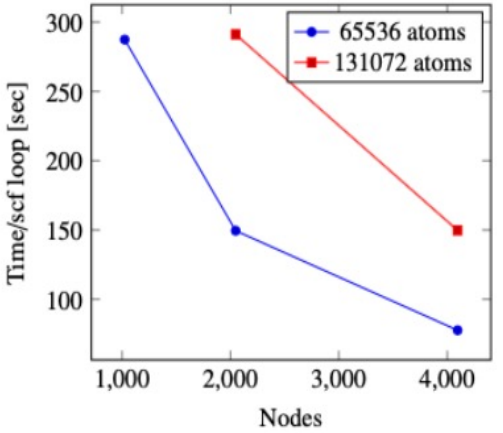
Future work

- Capabilities for non-metallic quantum materials
- Calculation of forces for ab-initio relaxation and first-principles molecular dynamics.



Moving from CUDA to HIP

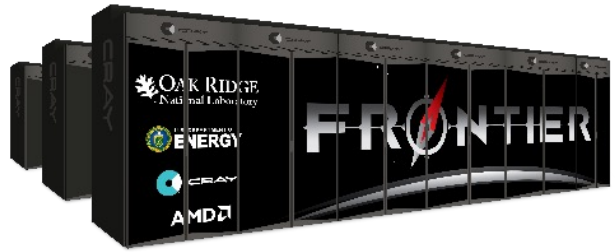


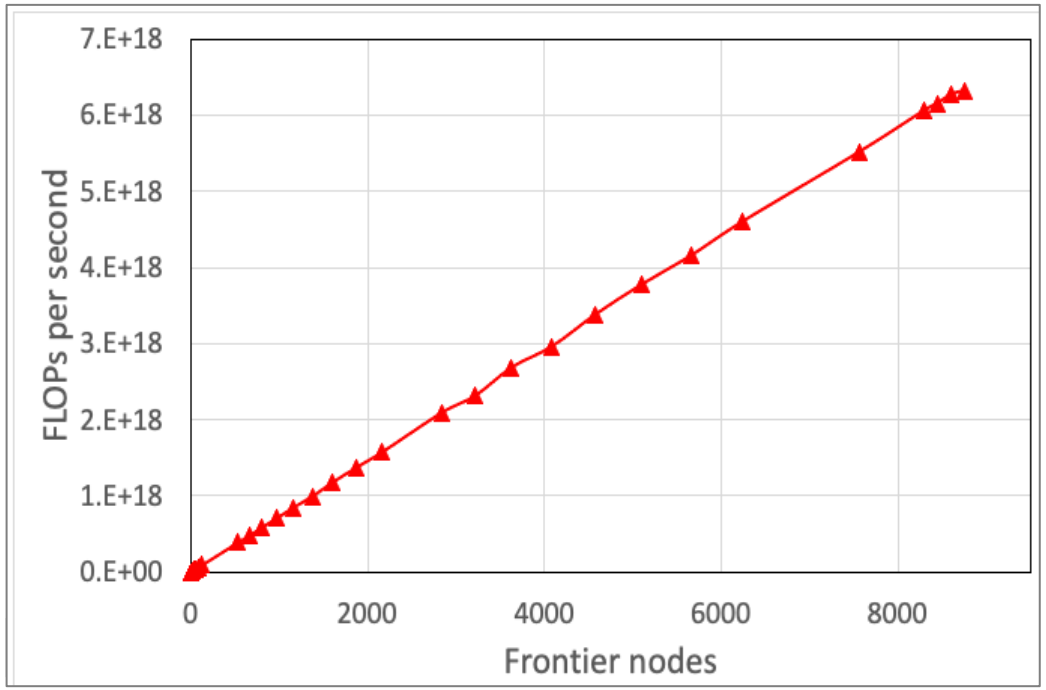


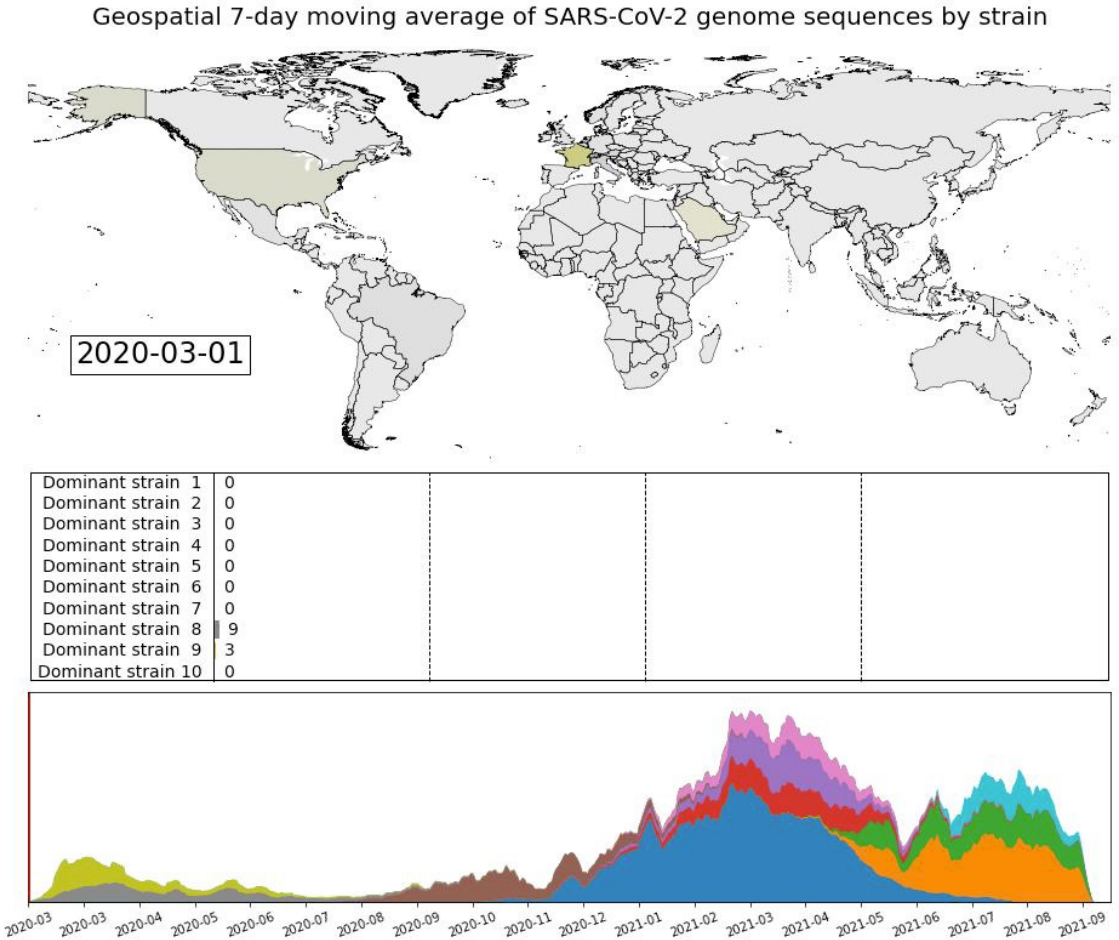
Weak (left) and strong (right) scaling results of LSMS for FePt calculations on Frontier

CoMet for correlation analysis on Frontier

- Comet is used to compute similarity metrics from large datasets in genomics, climate, and other fields
- Currently being used to analyze the geospatial and temporal evolution of SARS-CoV-2 variants
- CoMet has achieved up to **6.6 ExaFlops** mixed precision performance on Frontier (3-way DUO method)

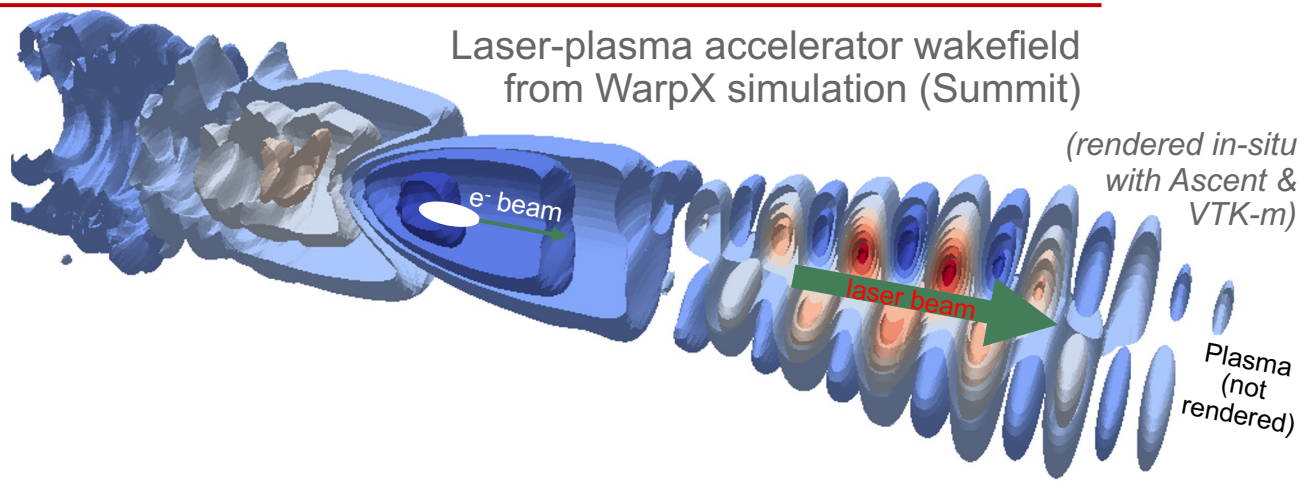






WarpX simulates plasma accelerators on thousands of Frontier nodes

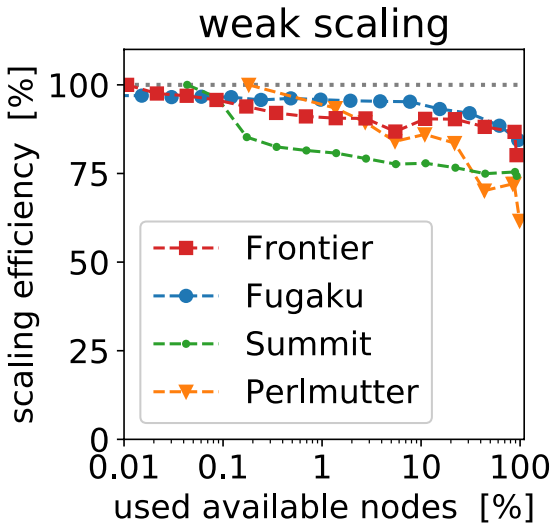
- **WarpX** is a Particle-In-Cell code developed by the Exascale Computing Project (ECP) for the modeling of plasma accelerators
  - Based on AMReX for CPU/GPU Mesh-Refinement
  - Portable CPU/GPU frameworks that avoid code duplication
  - Efficient data structures, memory & comms.



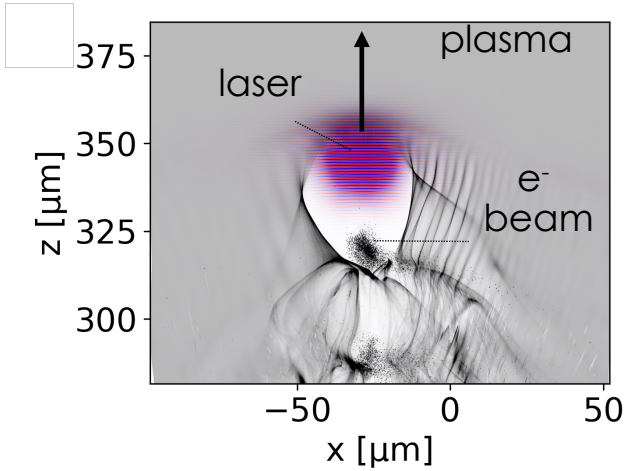
Laser-plasma accelerator wakefield from WarpX simulation (Summit)

(rendered in-situ with Ascent & VTK-m)

- **Large scale ECP Figure Of Merit run on 8,576 nodes**
  - Runtime: 100 timesteps w/ preloaded uniform plasma
  - $FOM_{Frontier}/FOM_{Edison}$~500



- **Large scale science runs performed on 2K & 8K Frontier nodes for Gordon Bell Prize submission**
  - Plasma acceleration of an electron beam generated by interaction of the laser with a plasma target
  - Demonstrated science runs at higher resolution than on Summit and Fugaku
  - WarpX team (LBNL+CEA Saclay+ENSTA+GENCI +Arm+ATOS+RIKEN) finalist of Gordon Bell 2022

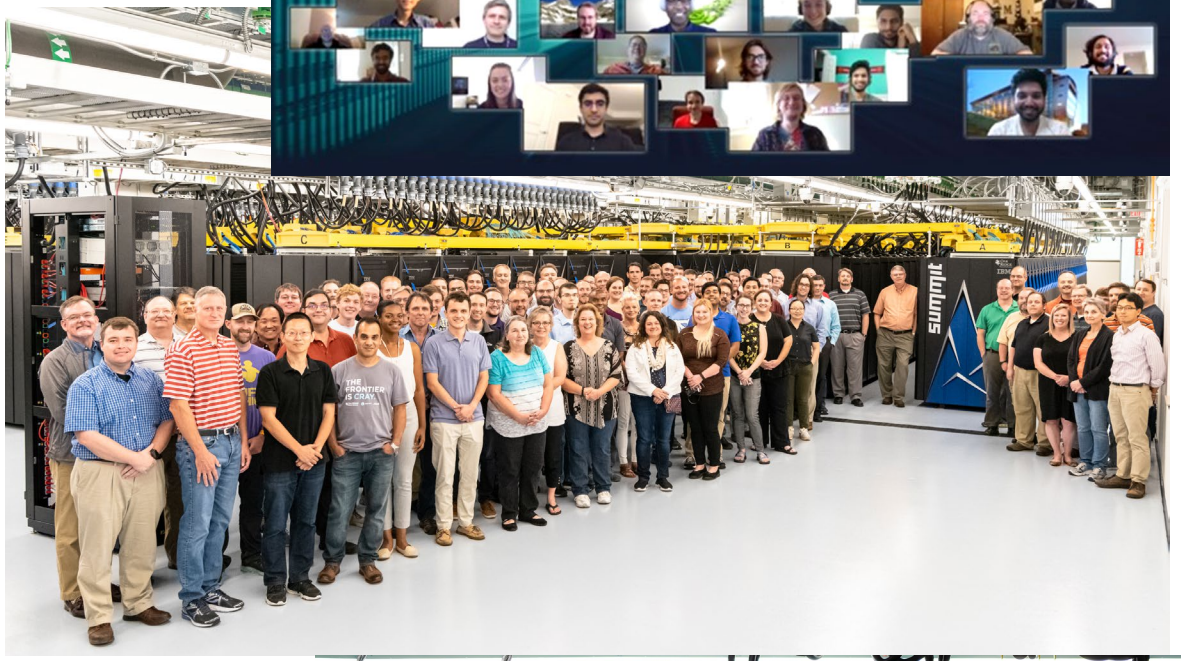


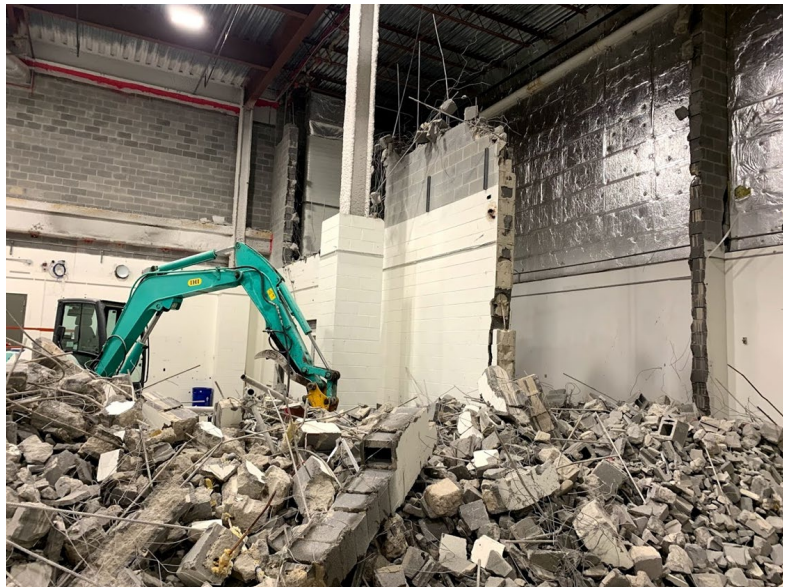
(figure from run on 4K Summit nodes)

(Frontier: 9 316 available nodes)

Many talented people helped make Frontier a reality

- Broad support from DOE HQ and Site Office
- 150 experts from 6 labs met in late 2018 to review technical proposals for Frontier
- 1,000 ECP staff
- 90 OLCF staff
- Over 100 electrical and mechanical workers
- Over 300 HPE and AMD engineers





Questions?


FRONTIER
OAK RIDGE
National Laboratory
DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA
U.S. DEPARTMENT OF
ENERGY