OAK RIDGE National Laboratory

Update on Exascale Systems - Frontier

Justin L. Whitt
Leadership Computing Facility
Oak Ridge National Laboratory

ASCAC
July 2022

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

U.S. DEPARTMENT OF ENERGY

Frontier Update in 4 Parts

- System status
- Facility status
- Overcoming Supply Chain Issues
- Early User Experiences on Frontier Hardware

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

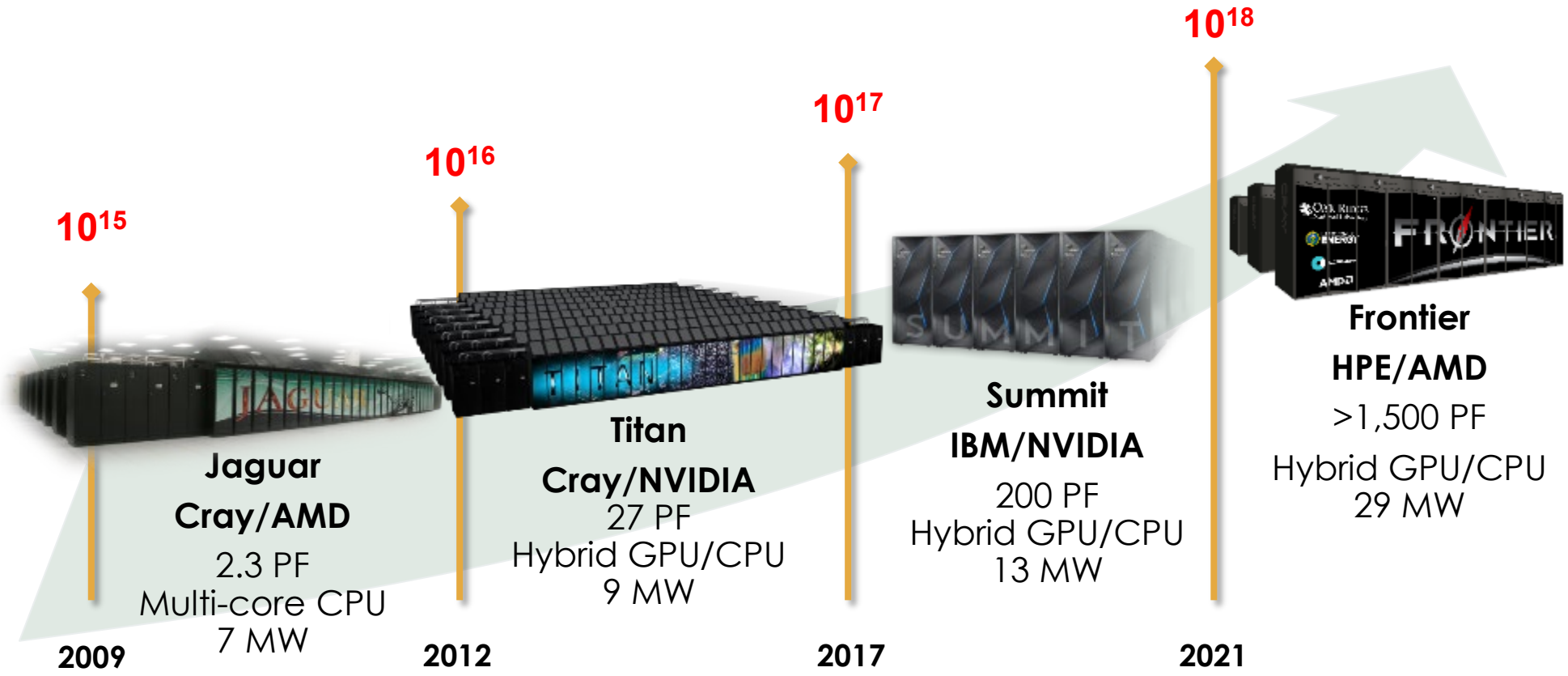
#1

6.88 exaflops on the HPL-AI benchmark.

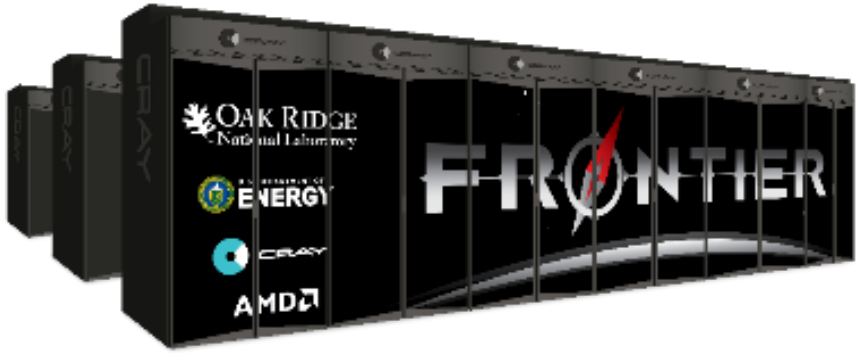
Sources: May 30, 2022 Top500 release

Oak Ridge Leadership Computing Facility – a DOE Office of Science User Facility

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.



Frontier System



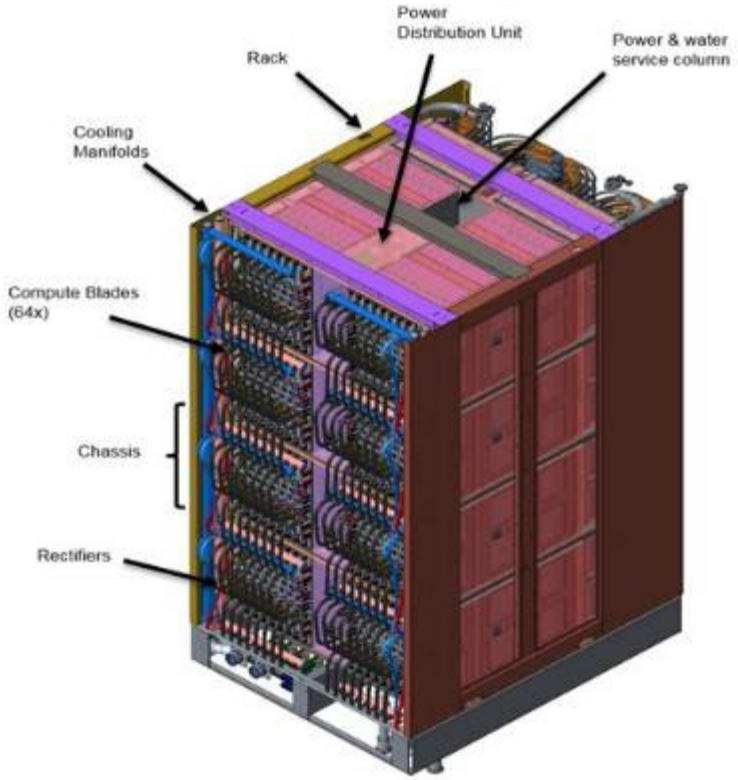
System

- 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

Frontier Cabinet

Olympus rack

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



Frontier Node

AMD extraordinary engineering

- 1 AMD "Trento" CPU (optimized Milan)
- 4 AMD MI250X GPUs

- 512 GiB DDR4 memory on CPU
- 512 GiB HBM2e total per node
- 4 Cassini NICs connected to the 4 GPUs

Compute blade

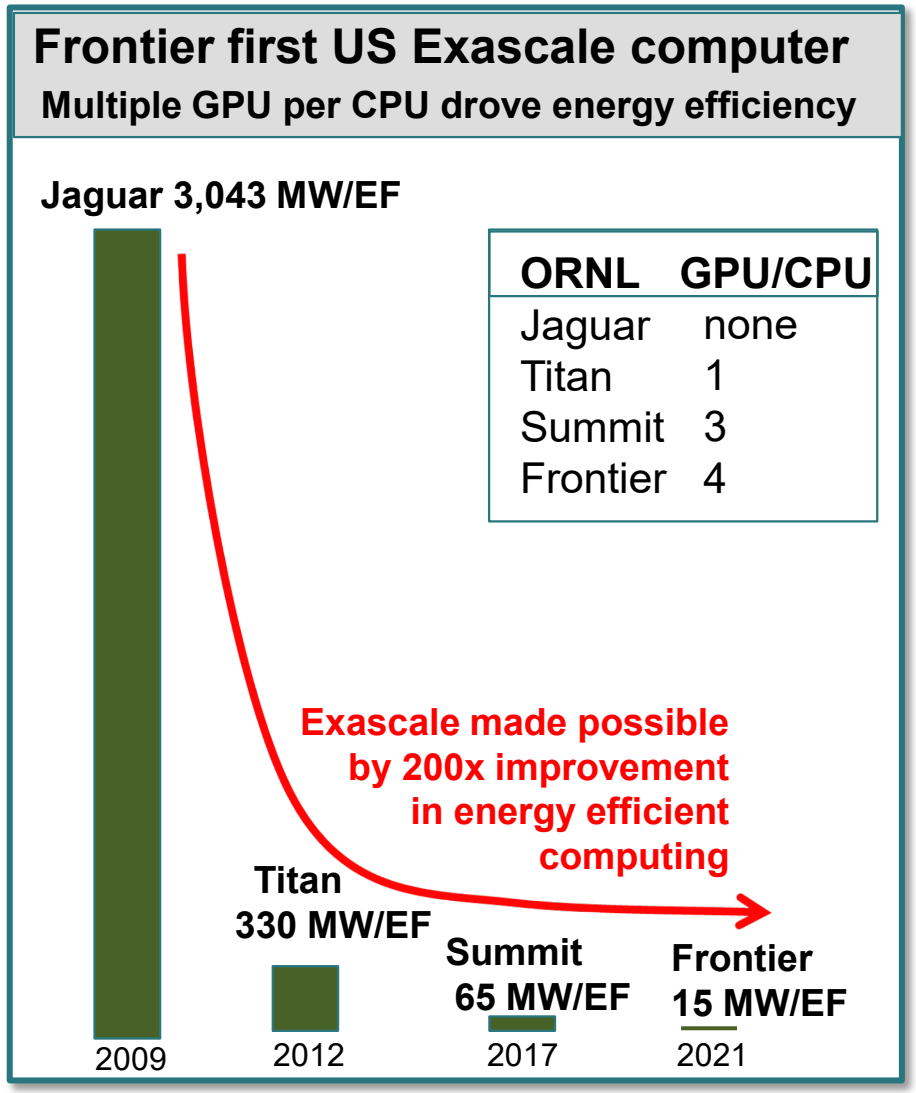
- 2 AMD nodes

All water cooled, even DIMMS and NICs

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



Facility Status

To provide the power, space, and cooling for Frontier

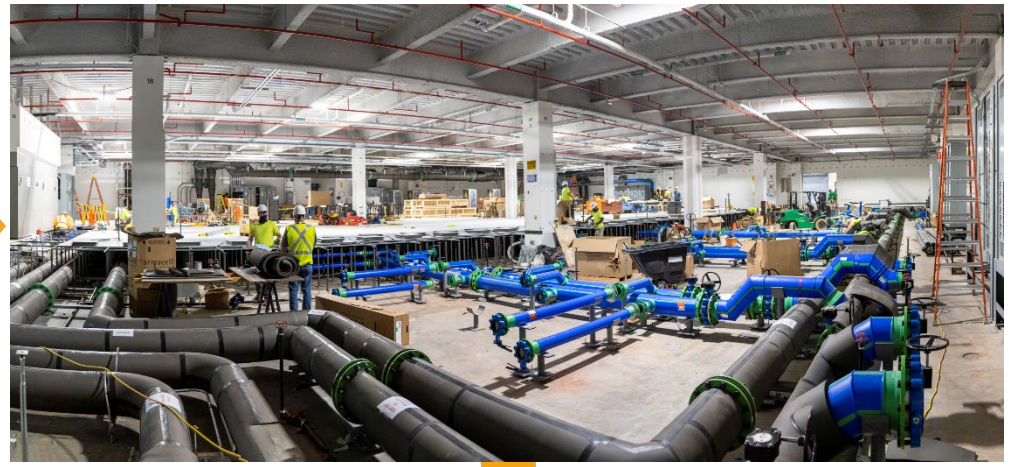
- 30 offices, 8 laboratories, and a 20,000 s.f. data center were repurposed

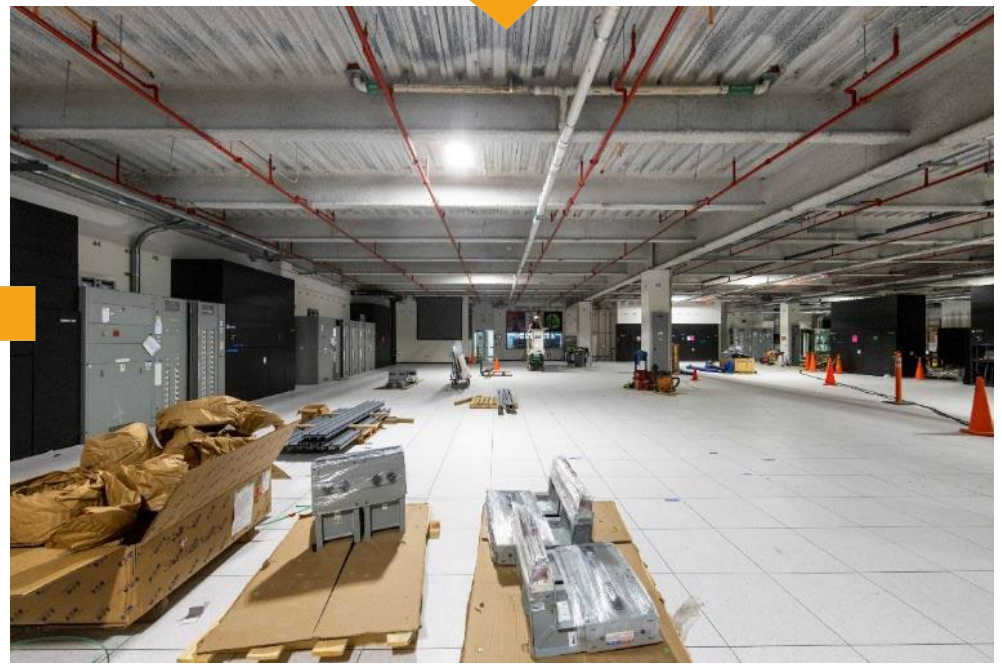
To provide 40 MW of cooling

Additional Cooling Towers

40 MW of power

The old Titan data center becomes the new Frontier data center





Overcoming Supply Chain Issues

By the end of 2020 the Part Shortage Had Hit in Earnest!

World-wide part shortages are a BIG problem if you are building an Exascale Computer. When HPE began ordering parts, many suppliers said the lead time on orders had increased an additional 6-12 months

60 Million parts needed for Frontier

685 Different part numbers used in Frontier

167 Frontier part numbers affected by shortages
(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 MI250 cards for Frontier

It wasn't just exotic parts like new CPUs and GPUs, but also parts such as voltage regulators, oscillators, power modules

Frontier Build– Supply Chain Remained a Constant Battle Delaying Final Delivery from Summer to Fall of 2021

HPE saw commitments for parts deliveries from sub contractors being broken weekly as the chip shortage got worse. Had to call every supplier every week (sometimes every day)

HPE and AMD had 15 people whose sole job was to try to find the needed parts or alternatives for Frontier. Using HPE's size to negotiate with suppliers, looking for handfuls of parts in warehouses or at other companies who were also stuck because of chip shortage.

April 30 – July 15: Initial shortage of 167 part numbers reduced down to 1 part number

- An oscillator needed for Slingshot blade
- July 15th only found enough to build 63 of 74 cabinets (still looking for about 8,000 more)
- It took three more weeks to find all 8,000

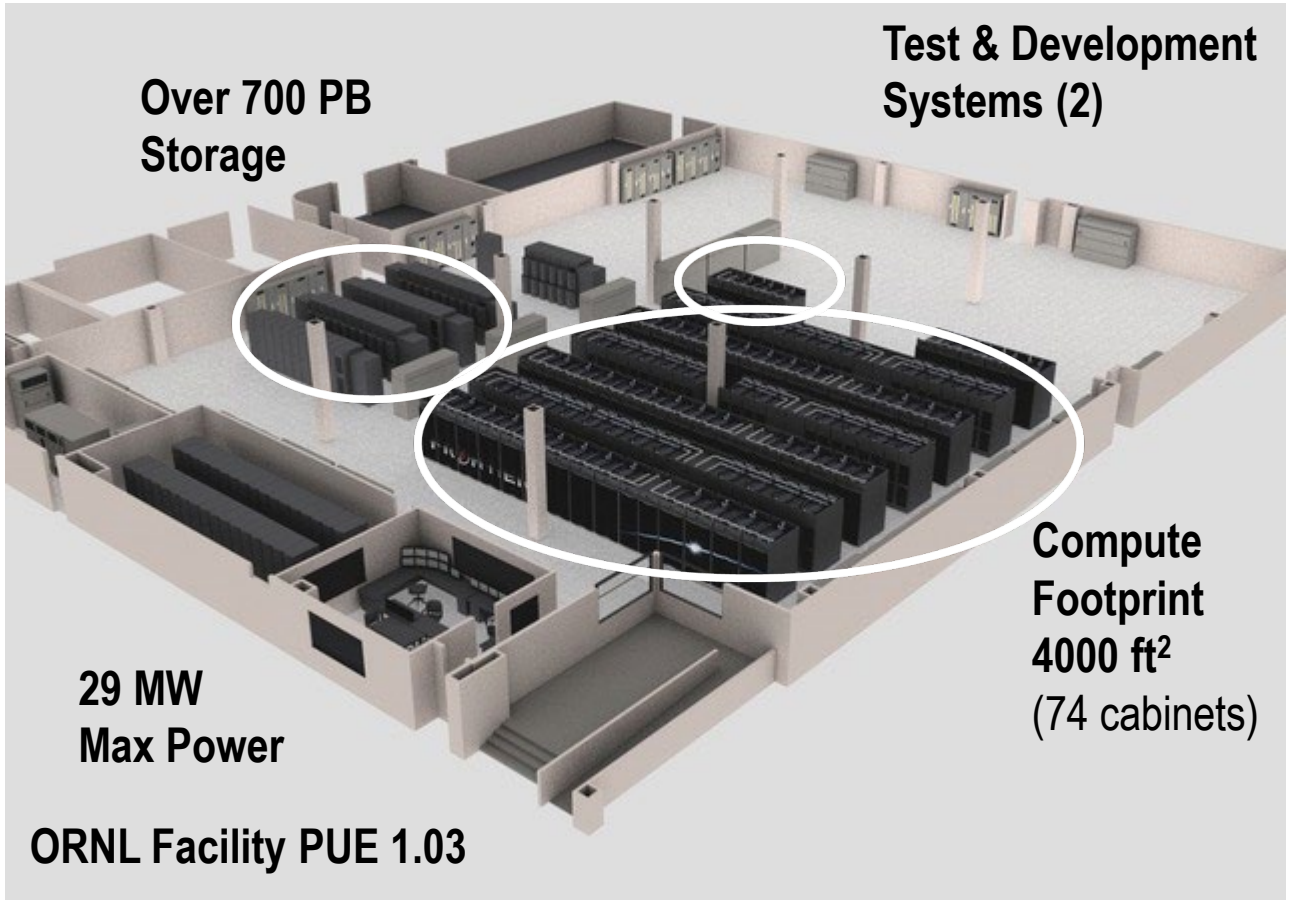
PCA Shortages	4/30	5/7	5/14	5/21	5/28	6/4	6/11	6/18	6/25	7/2	7/9	7/16
Critical Shortages	167	69	46	39	30	28	28	11	6	3	2	1
New Shortages	0	0	0	1	0	0	0	1		0	0	0
Total	167	69	46	40	30	28	28	12	6	3	2	1

The final parts arrived the morning the last Frontier node was assembled

Last Cabinet of Frontier Delivered to ORNL October 18, 2021

Thanks to Heroic Efforts of the HPE and AMD teams

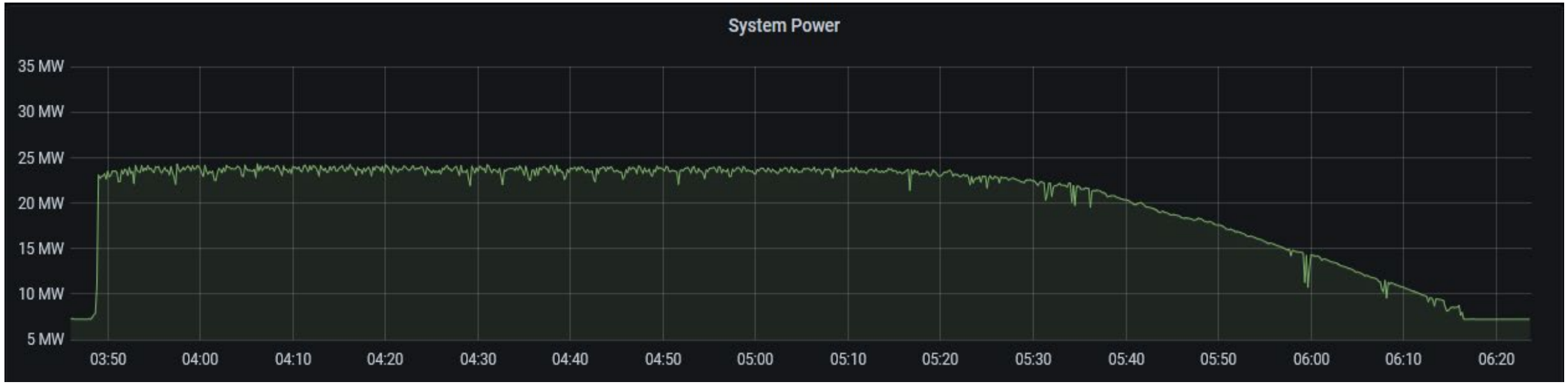
Last cabinet being rolled into place.
(Each cabinet weighs 8,000 lbs.)



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

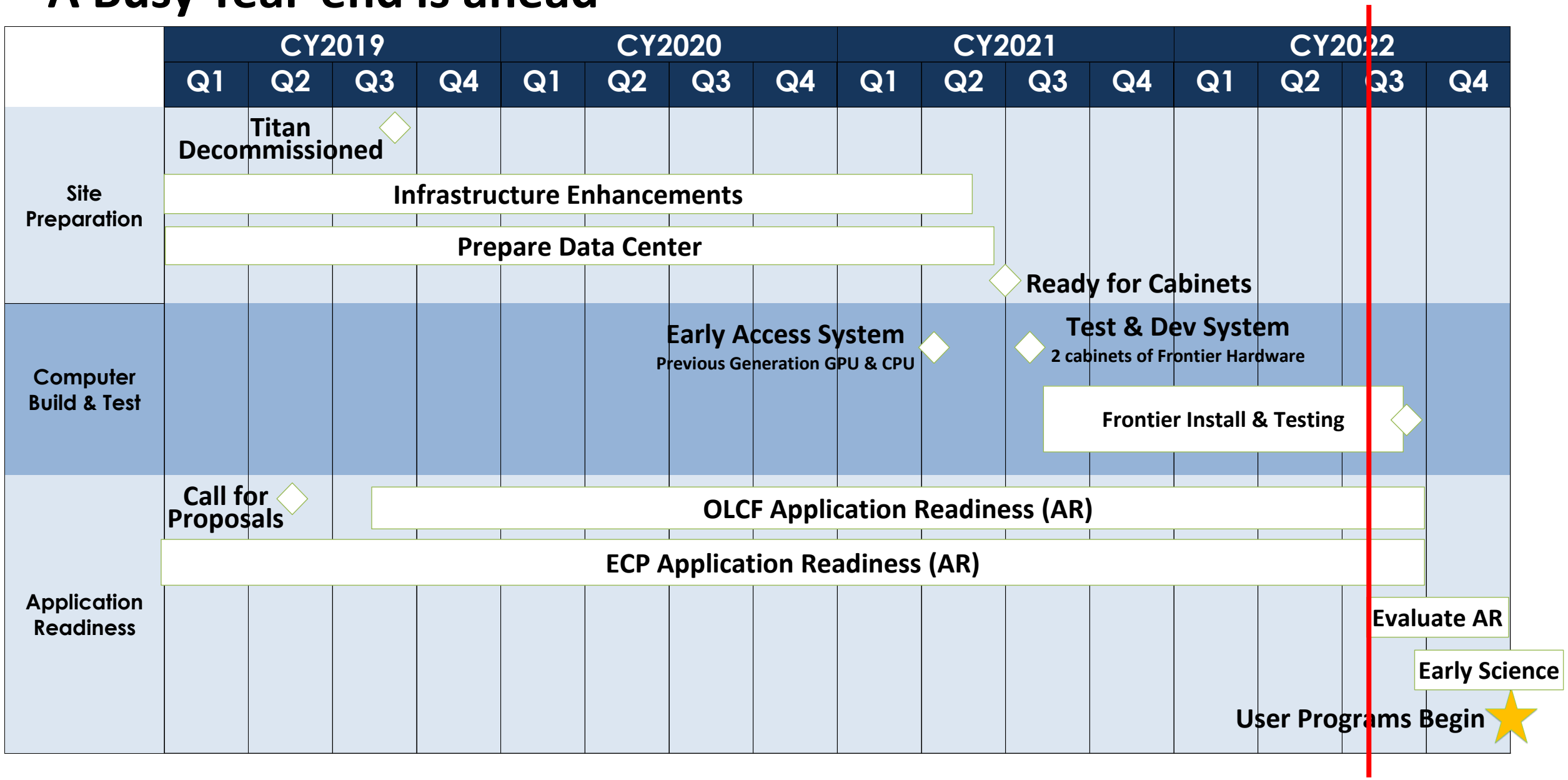
- We fell into a pattern of repairing hardware, updating software, and tuning the system by day
- And running benchmarks like HPL at night



- 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

A Busy Year-end is ahead



Early User Experiences on Frontier Hardware

Image by
Karsten Winegeart

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

OLCF-5 Applications		ECP Applications	
Astrophysics	CHOLLA	Astrophysics	ExaStar
Molecular Dynamics	NAMD	Astrophysics	ExaSky
		HEP	LatticeQCD
Materials Science	LSMS	Chemistry	NWCHEMeX
Biology/Health	CoMet	Chemistry	GAMESS
		Combustion	PELE
Fluid Dynamics	GESTS	Energy	ExaSMR
Nuclear Physics	NUCCOR	Energy	WDMApp
		Climate	E3SM
Plasma Physics	PIConGPU	Additive Manufacturing	ExaAM
Subsurface Flow	LBPM	Biology	ExaBiome
		Electric Grids	ExaSGD

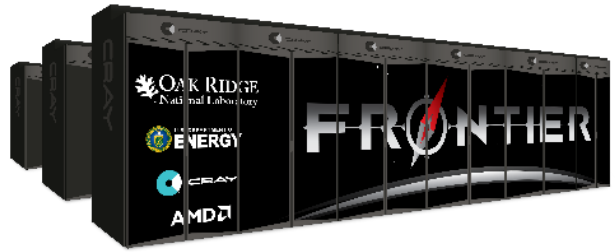
ECP Application Progress on Crusher

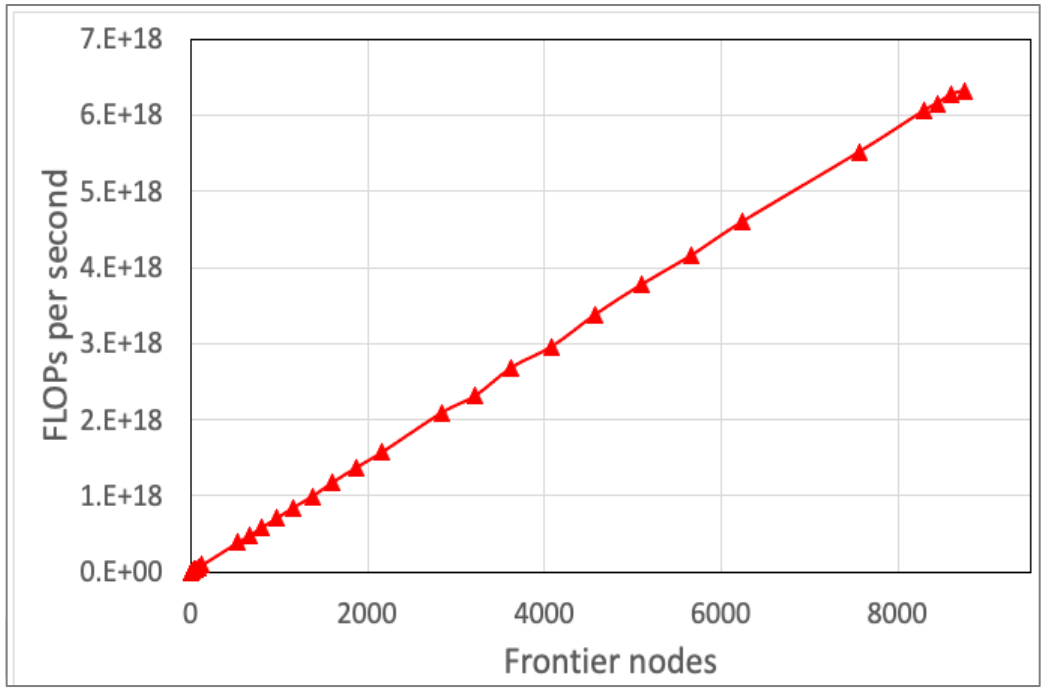
KPP-1 Application	Crusher (ORNL)
LatticeQCD	Ready
NWChemEx	Initial Build/Test
EXAALT	Improving Perf.
QMCPACK	Improving Perf.
ExaSMR	Ready
WDMApp	Improving Perf.
WarpX	Improving Perf.
ExaSky	Ready
EQSIM	Ready
E3SM-MMF	Improving Perf.
CANDLE	Ready

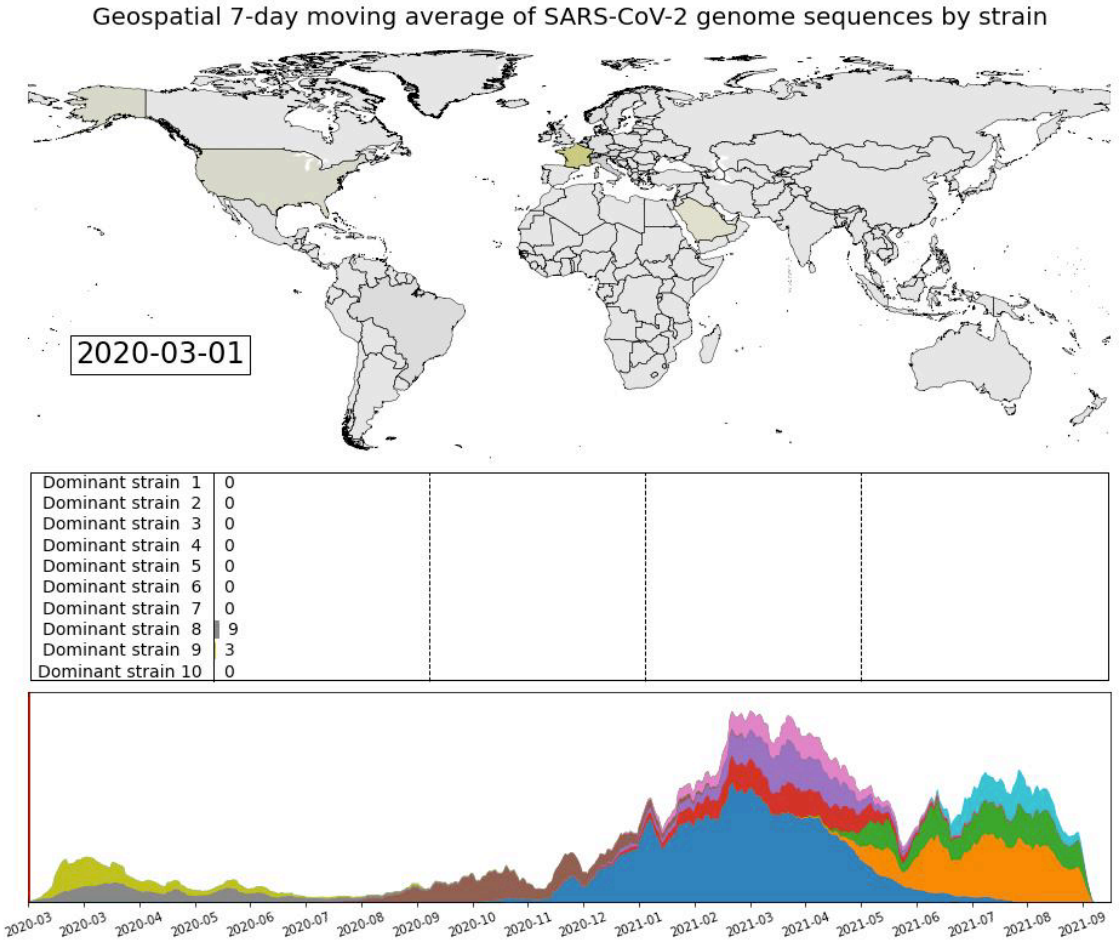
KPP-2 Application	Crusher (ORNL)
GAMESS	Blocked (ROCm 5.x)
ExaAM	Improving Perf.
ExaWind	Improving Perf.
Combustion	Improving Perf.
MFIX-Exa	Improving Perf.
ExaStar	Improving Perf.
Subsurface	Ready
ExaSGD	Improving Perf.
ExaBiome	Improving Perf (GasNet workaround)
ExaFEL	Blocked (Rocm 5.x)

CoMet for correlation analysis on Frontier

- Comet is used to compute similarity metrics from large datasets in genomics, climate and others
- 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)







Large Scale Density Functional Theory at the Exascale with LSMS

Workflows and high-performance computations to predict materials properties

Research Topics

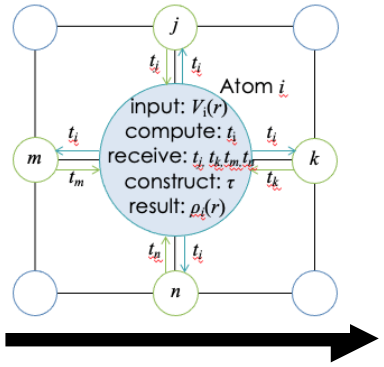
- Understanding the role of disorder and defects in materials for electronic and mechanical properties
- Complex magnetic order – topological magnetic structures and magnetism beyond ideal crystal

Recent Highlights

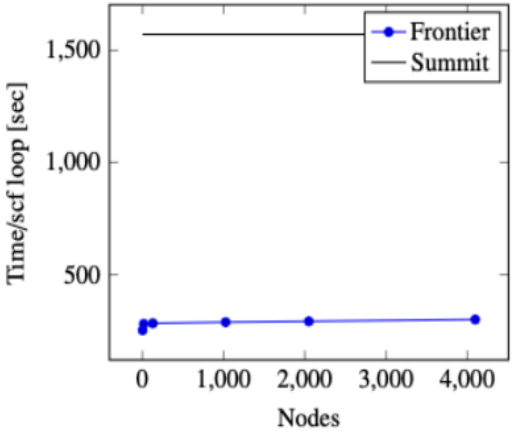
- Scaling of first principles calculations to O(100,000) up to O(1,000,000) atoms for the first time.
- Demonstrated scaling of LSMS on Frontier up to 1,048,576 atom FePt system on 8192 Frontier nodes.
- Speedup of LSMS from Summit to Frontier from combined hardware and software improvements is ~8x

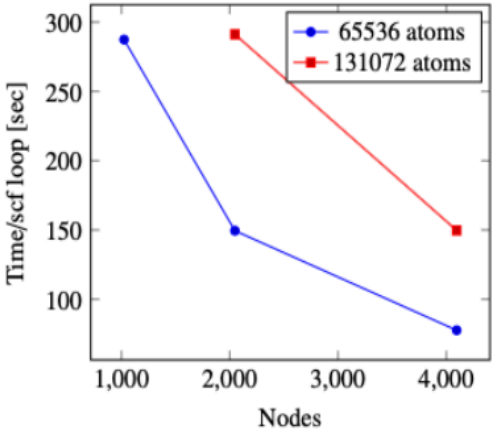
Future work

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



Moving from CUDA to HIP

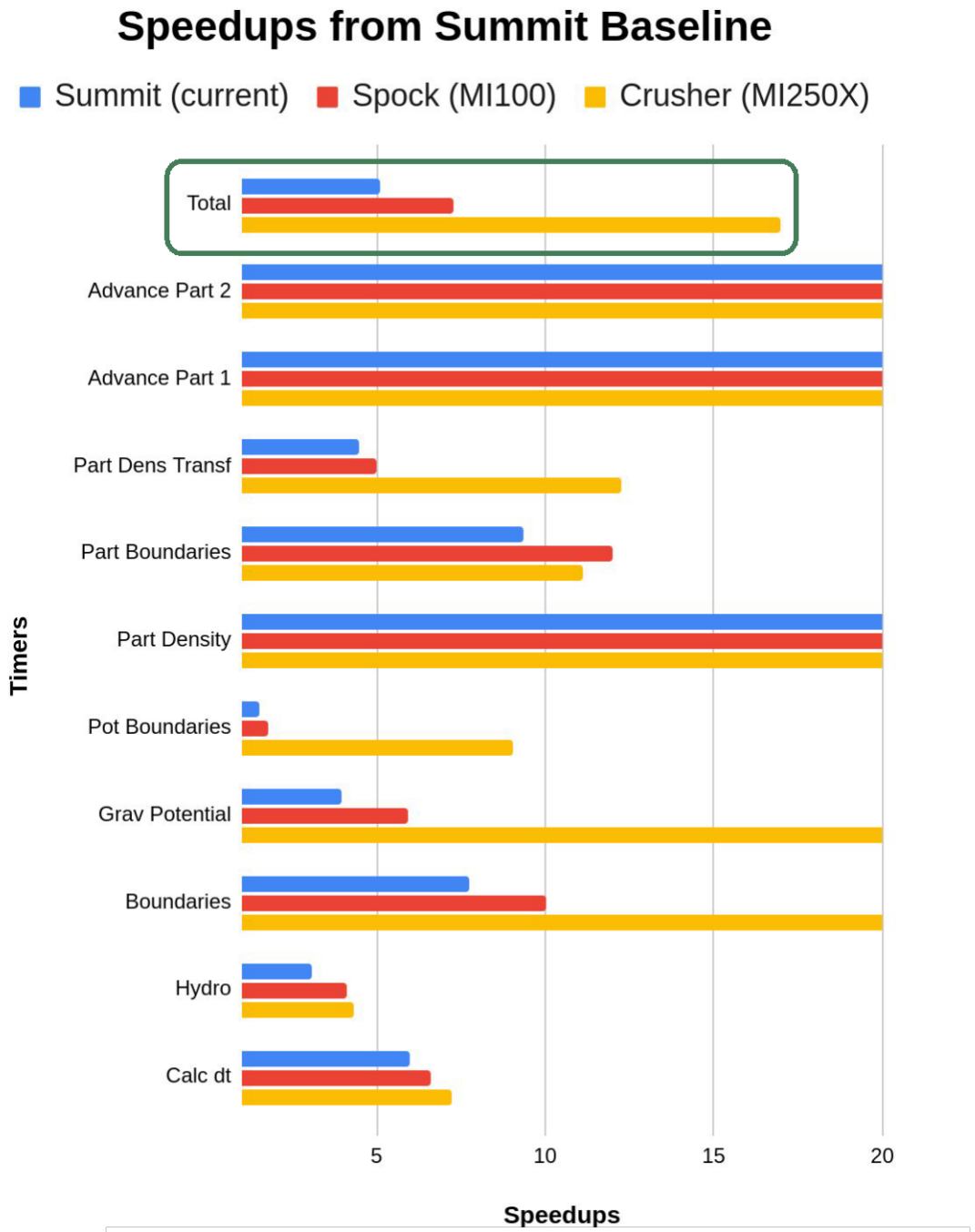




Weak (left) and strong (right) scaling results of LSMS for FePt calculations on Frontier

CAAR Cholla Status (Feb 2022)

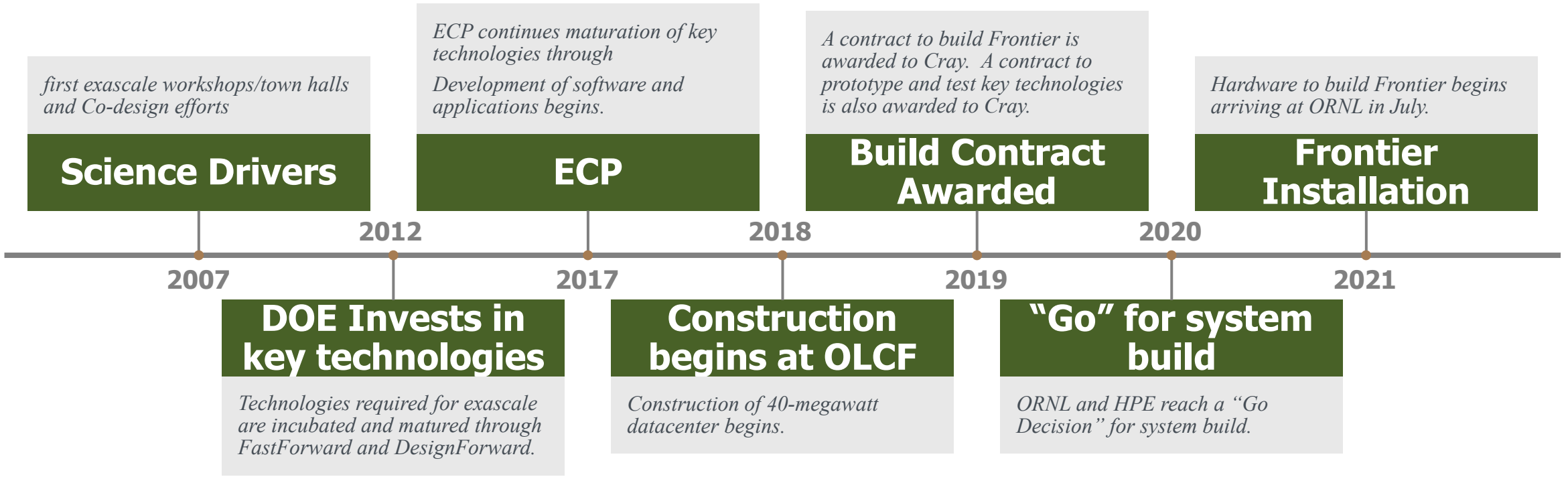
- Total speedups: **~16X** on 64 GPUs on Frontier hardware (Crusher) from baseline (see plot).
- Software development contributed to ~5X speedups on Summit (blue bars on the plot). Major highlights:
  - Made hydro grid fully GPU resident
  - Exploited GPU-aware MPI
  - Ported gravity solver to GPU
  - Ported particle solver to GPU
- Hardware improvements from Summit to Crusher: ~3X speedups
- Pending: Scaling up to the full Frontier



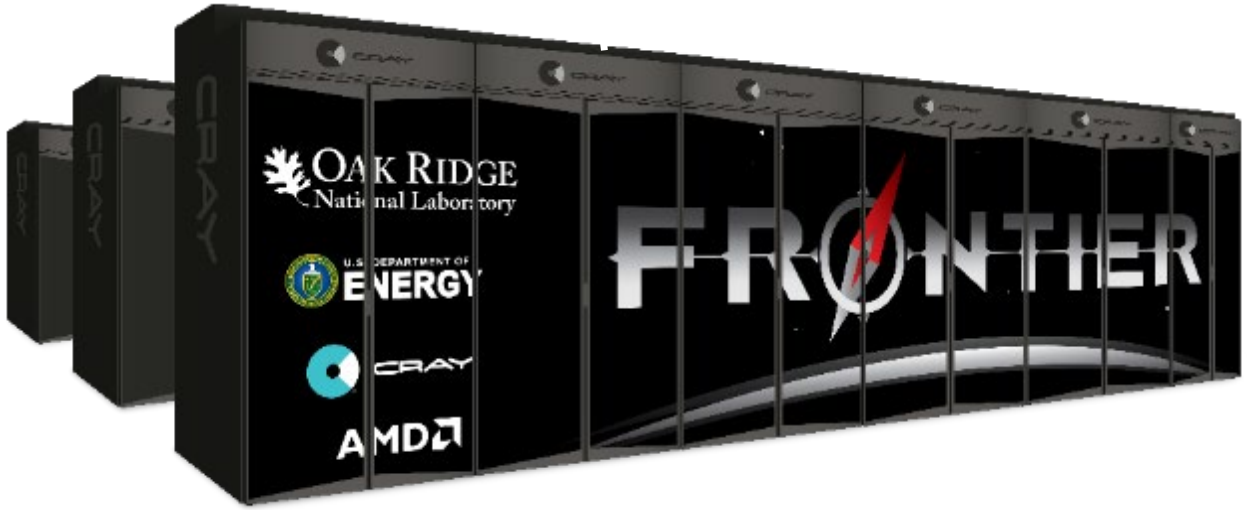
Many talented people helped make Frontier a reality

- 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

Decadal effort to deliver U. S. Exascale systems led to Frontier



Thank You


CRAY
OAK RIDGE
National Laboratory
ENERGY
CRAY
AMD
FRONTIER