OAK RIDGE National Laboratory

# UPDATE ON FRONTIER EXASCALE SYSTEM AND EARLY SCIENCE

Al Geist
Leadership Computing Facility
Oak Ridge National Laboratory

ASCAC Meeting
March 29-30, 2022

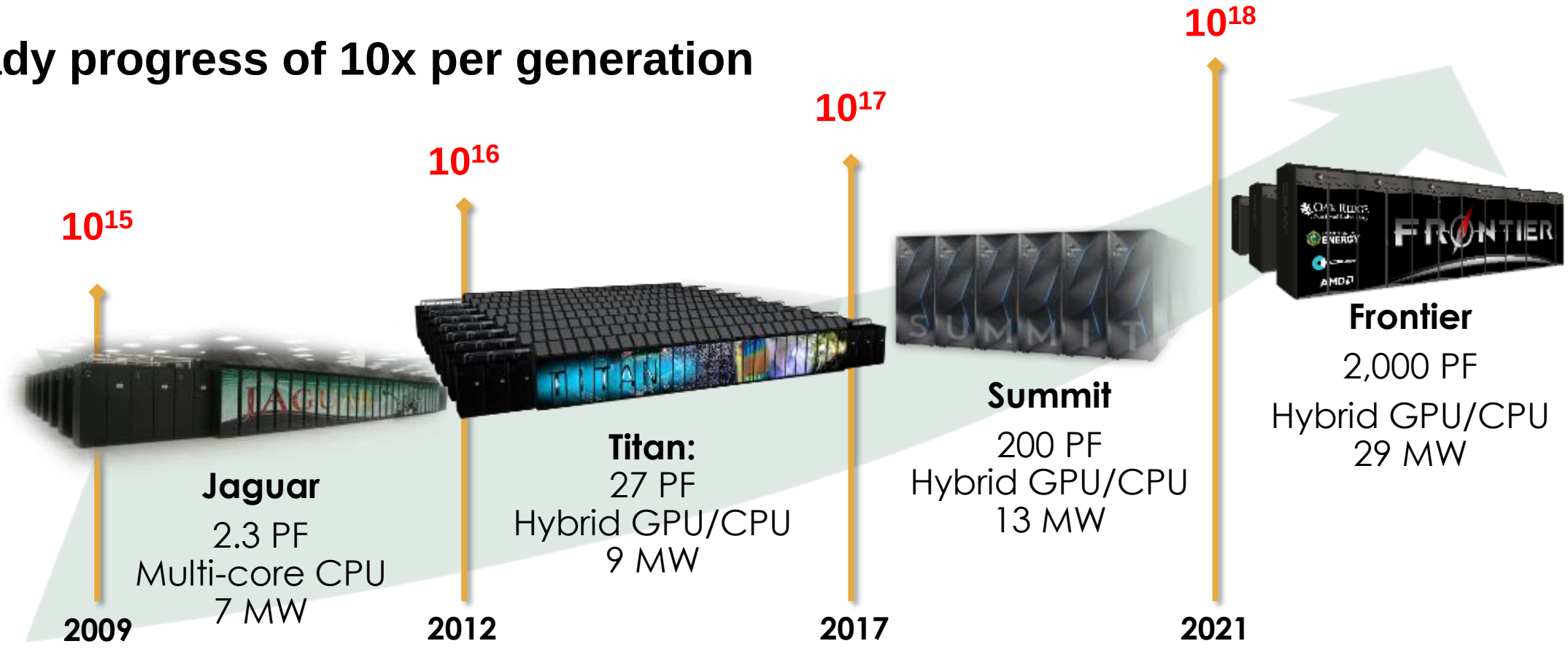
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

# Oak Ridge National Laboratory's Journey 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. |
|---|---|

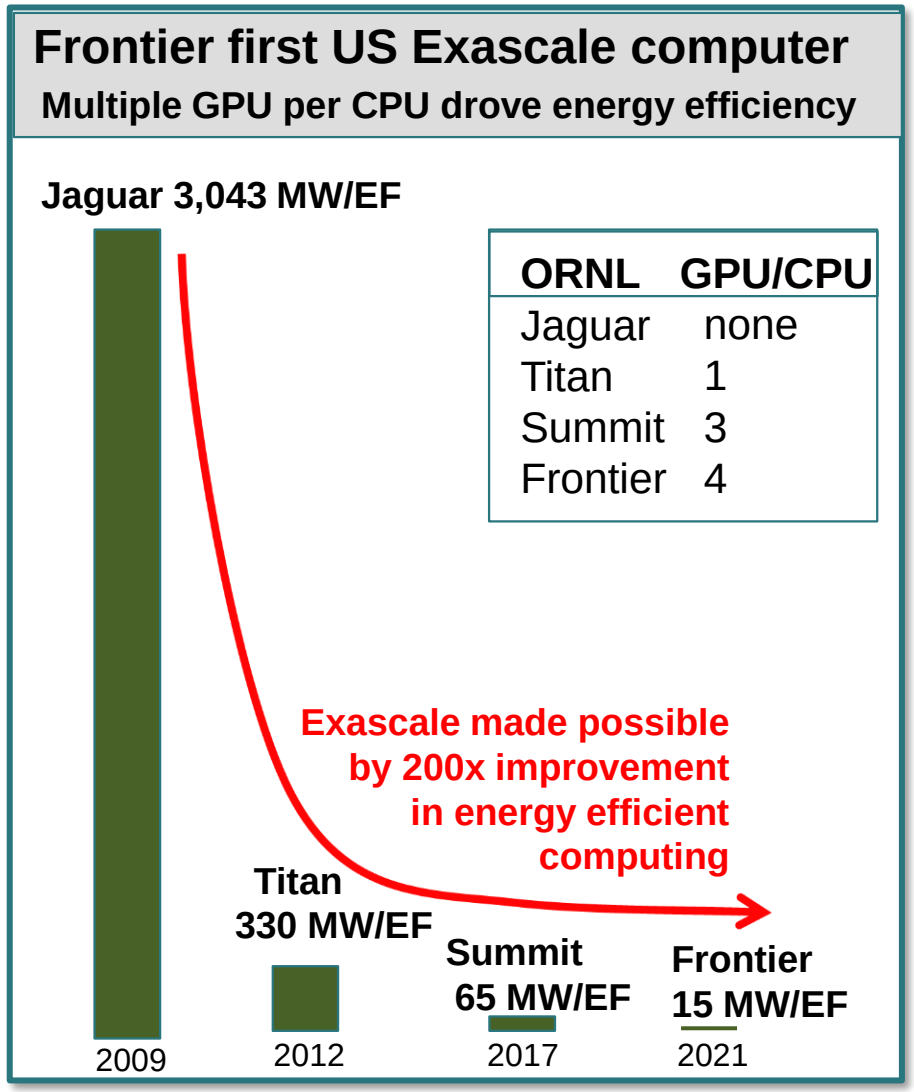
## Steady progress of 10x per generation



# 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**



# On our journey to Exascale, we found an architecture that could excel at simulation, data analytics, and artificial intelligence

As supercomputers got larger and larger, we expected them to be more specialized and limited to just a small number of applications that can exploit their growing scale

**We found that "Summit" architecture with few, large-memory, multi-GPU nodes excels at:**

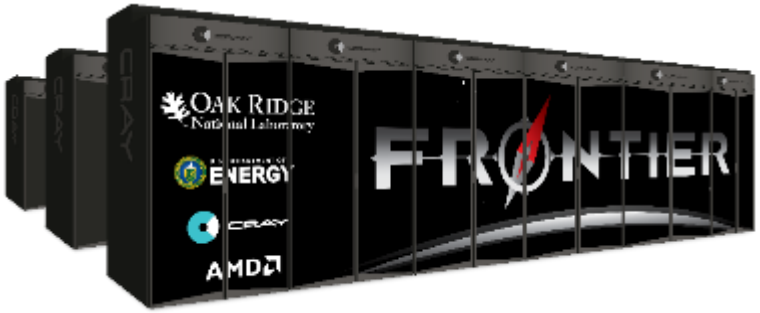

- Data analytics – CoMet bioinformatics application for comparative genomics.
  **Has achieved 2.36 ExaOps** mixed precision (FP16-FP32) on Summit (2018 Gordon Bell Winner)
- Deep Learning – Climate: neural network learns to detecting extreme global weather patterns
  **Has achieved a sustained throughput of 1.0 ExaOps** (FP16) on Summit

Frontier Exascale computer uses and improves on Summit's successful architecture

- 5 TB of on-node memory, 4 GPU per node, Peak of >10 ExaOps (FP16)

# Frontier Overview

## Extraordinary Engineering



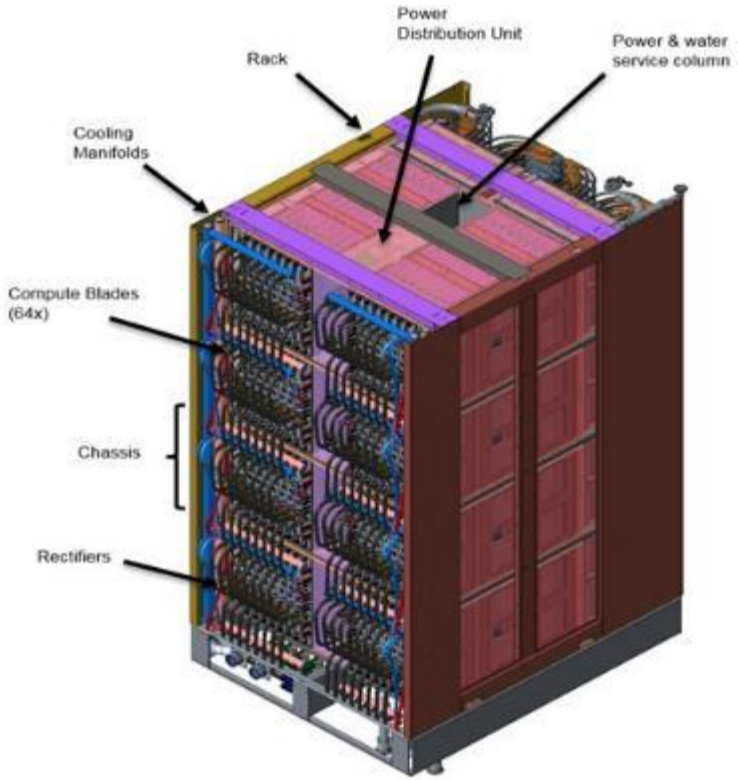

### System

- 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**

# Frontier multi-tier storage system is designed to excel at Data Science and AI for Scientific Discovery

| | Capacity | Performance | |
|---|---|---|---|
| **Multi-tier I/O Subsystem** | | **Read** | **Write** |
| | 37 PB Node Local Storage | 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 | |

Two 2TB SSD NVM per node Local Storage (Flash)

Gazelle SSD Storage board (Performance Tier and Metadata)

Moose HDD Storage board (Capacity Tier)

# 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.

**ORNL worked with ASCR to get DPAS rating for Frontier that helped prioritize USA part orders (DPAS was extended to Aurora and El Capitan)**

**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

# Supply Chain Remained a Constant Battle till Delivery

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 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

- July 15th only found enough to build 63 of 74 cabinets (looking for about 8,000 more)
- Took three more weeks to find all 8,000
- By that time had a couple more decommits on another part.

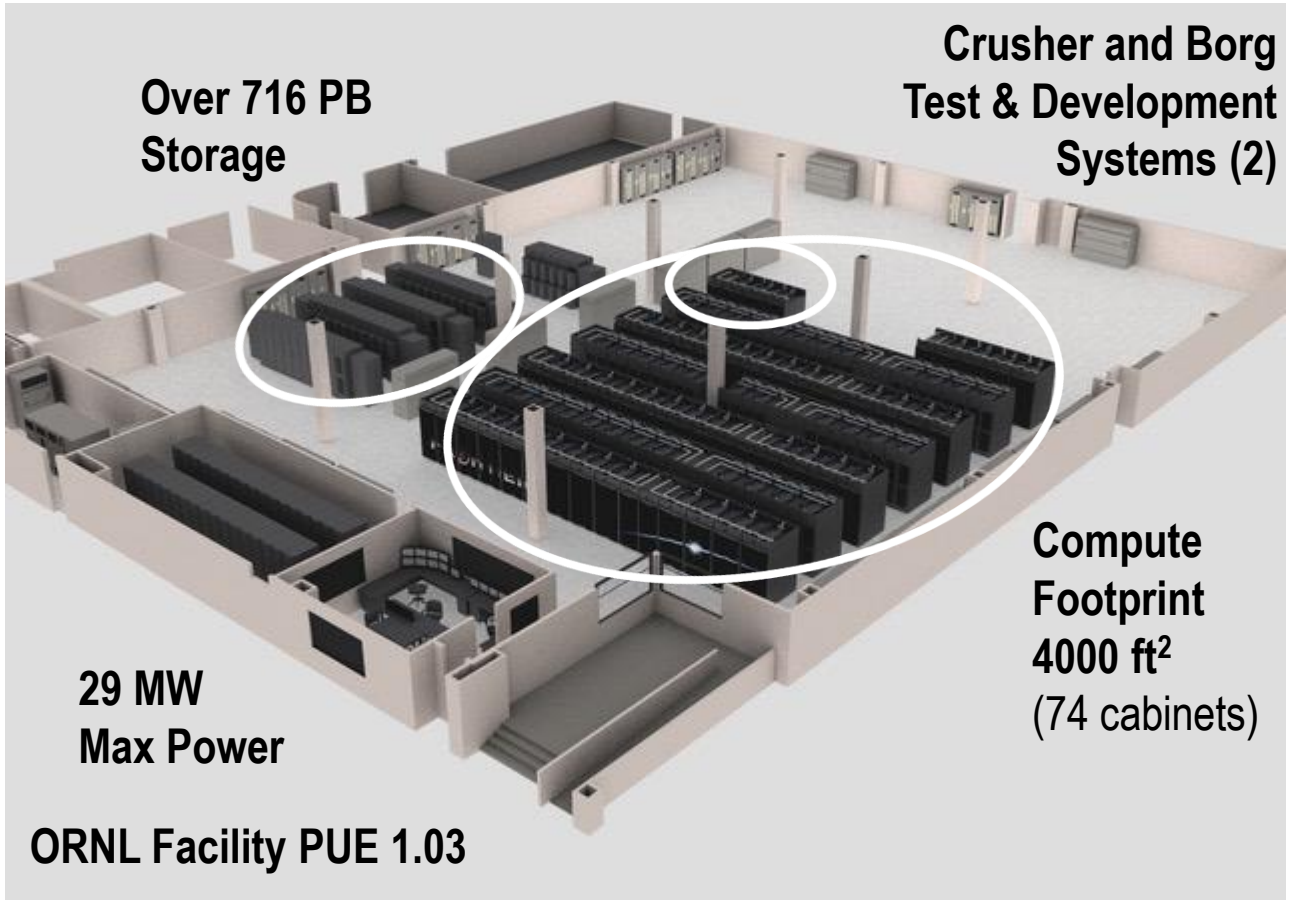
| PCA Shortages | 4/30 | 5/7 | 5/14 | 5/21 | 5/28 | 6/4 | 6/11 | 6/18 | 6/25 | 7/2 | 7/9 | 7/16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Critical Shortages | 167 | 69 | 46 | 39 | 30 | 28 | 28 | 11 | 6 | 3 | 2 | 1 |
| New Shortages | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | | 0 | 0 | 0 |
| Total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

## 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

# Getting Frontier Ready for Early Science

- As we saw with Titan and Summit, it takes a number of months to get all the hardware and software stabilized
- HPE continues testing and stabilization of Frontier and its file system
- Early Science Teams in CAAR and ECP got access to the “Crusher” Test & Development system in November 2021.
- Rest of ECP users (~800) given Crusher access January 2022.

ECP is scheduled for full Frontier access July 2022
INCITE use of Frontier scheduled for January 2023

**“Crusher” TDS system**

- 2 cabinets of Frontier HW
- 192 nodes
- Slingshot 11 w/ Cassini
- Same Software as Frontier

# Getting Users Ready for Early Science

## Crusher Training  January 13, 2022

- HPE and AMD presented Frontier architecture details, programming environment features and tips and tricks, ORNL provided login instructions and a *Crusher Quick Start Guide*

## Two Crusher Hackathons in February for CAAR and ECP Early Science Teams

| February 9-11, 2022 |
|---|
| NAMD |
| LSMS |
| CoMet |
| GETS |
| NuCCOR |
| PIConGPU |
| LBPM |
| ExaBiome |
| FUN3D |

| February 15-17, 2022 |
|---|
| ExaStar |
| LatticeQCD |
| NWCHEMex |
| GAMESS |
| PELE |
| ExaSMR |
| WDMApp |
| E3SM |
| ExaAM |

**Quotes from Hackathon**

"Great interactions with HPE and AMD staff in resolving issues"

"Reduced unit runtime from 8 hr. to 11 min."

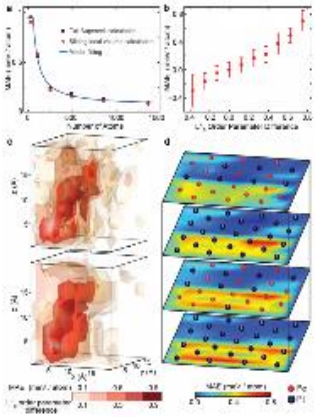
"Quick help and  learned some useful tricks and tips"

"Got 4x speedup on ExaAM PicassoMPM code"

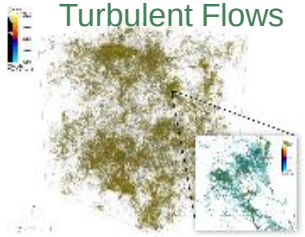
# Initial CAAR Early Science Results on Crusher

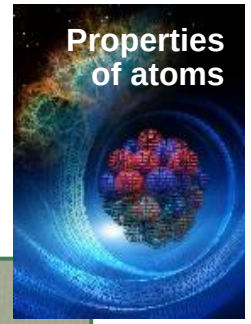
| Science Area | CAAR App | Recent Results on Crusher |
|---|---|---|
| **Advanced materials** | **LSMS** | MI250 getting up to 10x speedup over Summit V100 |
| **Turbulent Flows** | **GESTS** | Crusher GCD* achieves 6x speedup over Summit V100 |
| **Porus Media** | **LBPM** | Crusher GCD slightly faster than Summit V100. |
| **Plasma Physics** | **PIConGPU** | Seeing 2.5x – 5x speedup over Summit |
| **Atomic nucleus** | **NuCCOR** | Crusher MI250 performance gains of up to 8x over Summit V100 |
| **Health** | **CoMet** | Has been run on Frontier up to 3,210 nodes |
| **Astrophysics** | **Cholla** | Total of 15x speedup = Crusher HW getting additional 3x over Summit + 5x from SW |

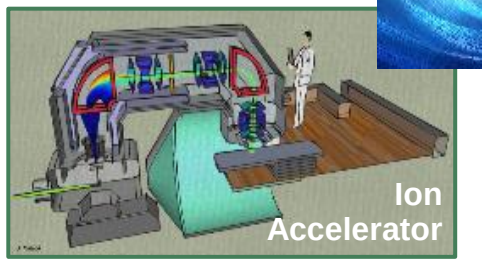
Magnetic anisotropy in FePt nanoparticle



Turbulent Flows







Astrophysics hydrodynamics

* MI250 GPU is composed of two GCDs

# Progress on Crusher by ECP KPP-1 Applications

## Apps selected to demonstrate performance improvement for mission-critical problems

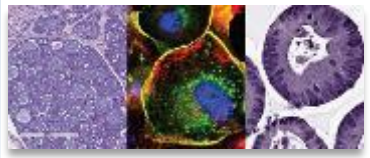
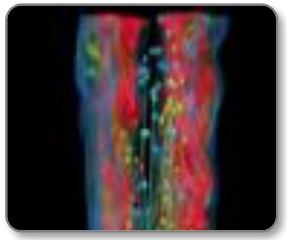
| Science Area | ECP Project | Crusher (TDS) |
|---|---|---|
| **Quantum Chromodynamics** | **LatticeQCD** | Improving Performance |
| **Chemistry (Biofuels)** | **NWChemEx** | Initial Build/Test |
| **Extreme Materials (MD)** | **EXAALT** | Improving Performance |
| **Quantum Materials (QMC)** | **QMCPACK** | Blocked (MPI) |
| **Nuclear Reactors (SMRs)** | **ExaSMR** | Improving Performance |
| **Fusion Plasmas** | **WDMApp** | Improving Performance |
| **Particle accelerators** | **WarpX** | Improving Performance |
| **Cosmology** | **ExaSky** | Improving Performance |
| **Earthquakes** | **EQSIM** | Improving Performance |
| **Climate Change** | **E3SM-MMF** | Improving Performance |
| **Cancer Research** | **CANDLE** | Improving Performance |

# Progress on Crusher by ECP KPP-2 Applications

## Apps selected to broaden the reach of exascale science and mission capability

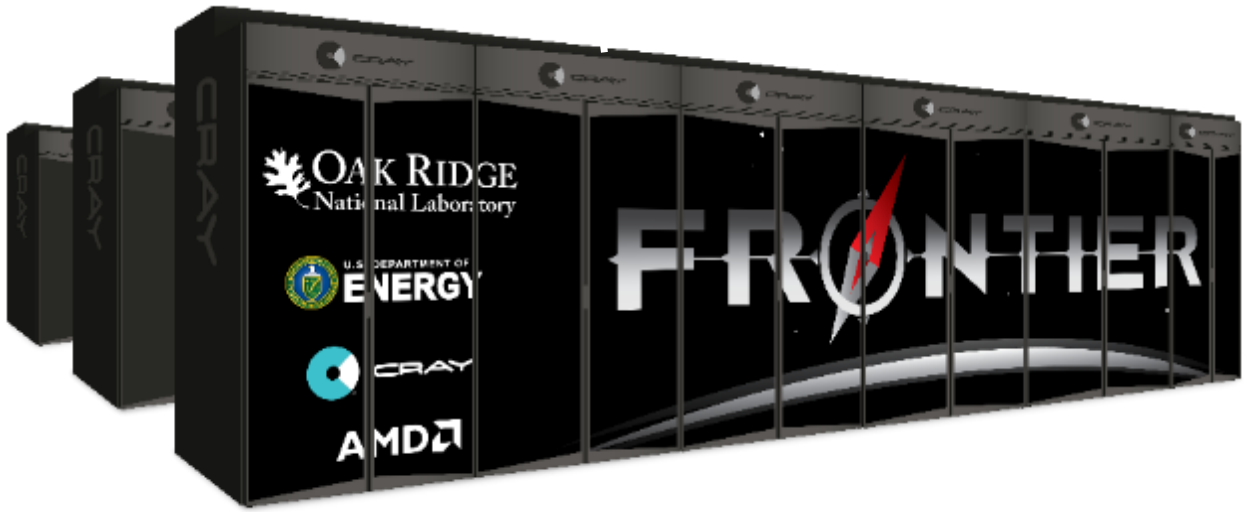
| Science Area | ECP Project | Crusher (TDS) |
|---|---|---|
| **Catalysis** | **GAMESS** | Blocked (ROCm 5.x) |
| **Additive Manufacturing** | **ExaAM** | Improving Performance |
| **Wind Energy** | **ExaWind** | Improving Performance |
| **Combustion** | **PELE** | Improving Performance |
| **Carbon Capture** | **MFIX-Exa** | Improving Performance |
| **Astrophysics** | **ExaStar** | Improving Performance |
| **Subsurface** | **Subsurface** | Improving Performance |
| **Energy Grid** | **ExaSGD** | Improving Performance |
| **Metagenomics** | **ExaBiome** | Blocked (GASNet) |
| **LCLS Molecular Structure** | **ExaFEL** | Improving Performance |

# ECP Application Portfolio – Early Science runs on Frontier

| Earth system | Energy security | Health care | Scientific discovery | Economic security |
|---|---|---|---|---|
| **Climate Change** | **Reliable and efficient planning of the power grid** | **Accelerate and translate cancer research** (partnership with NIH) | Cosmological probe of the standard model of particle physics | **Additive manufacturing** of qualifiable metal parts |
| Subsurface use for **carbon capture**, petroleum extraction, waste disposal | Turbine **wind plant** efficiency | Developing AI for Precision Drug Therapy in Fight Against Cancer | Validate fundamental laws of nature | Scale up of **clean fossil fuel** combustion |
| Accurate regional impact assessments in **Earth system models** | Design and commercialization of **Small Modular Reactors** |  | Find, predict, and control materials and properties | **Biofuel** catalyst design |
| Stress-resistant crop analysis and catalytic conversion of **biomass-derived alcohols** | Nuclear fission and fusion reactor **materials design** | | Light source-enabled analysis of protein and molecular structure and design | **Seismic** hazard risk assessment |
| **Metagenomics** for analysis of biogeochemical cycles, climate change, environmental remediation | High-efficiency, low-emission **combustion engine** and gas turbine design | | Predict and control magnetically confined fusion plasmas |  |
| |  | | Demystify origin of chemical elements | |

**Once KPP goal is achieved, ECP team can graduate to Early Science allocation on Frontier and begin using their ECP codes for Early science results**

# Questions?


OAK RIDGE
National Laboratory
U.S. DEPARTMENT OF
ENERGY
CRAY
AMD
FRONTIER