Oak Ridge National Laboratory

ORNL Site Report & Feature Discussion

Matt Ezell and Paul Peltz

SLUG 2024
September 12, 2024
University of Oslo, Norway

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

U.S. DEPARTMENT OF ENERGY

ORNL facts and figures

>$1B
modernization investment

$2.78B
FY23 budget authorization

>6,900
employees

Nation's most diverse energy portfolio

3,600
research guests annually

Nation's largest materials research portfolio

World-class research reactor

World's most intense neutron source

Managing major DOE projects: US ITER, exascale computing

8
DOE user facilities

>2,600
journal articles published in FY23

#1
fastest computer in the world

291
invention disclosures in FY23

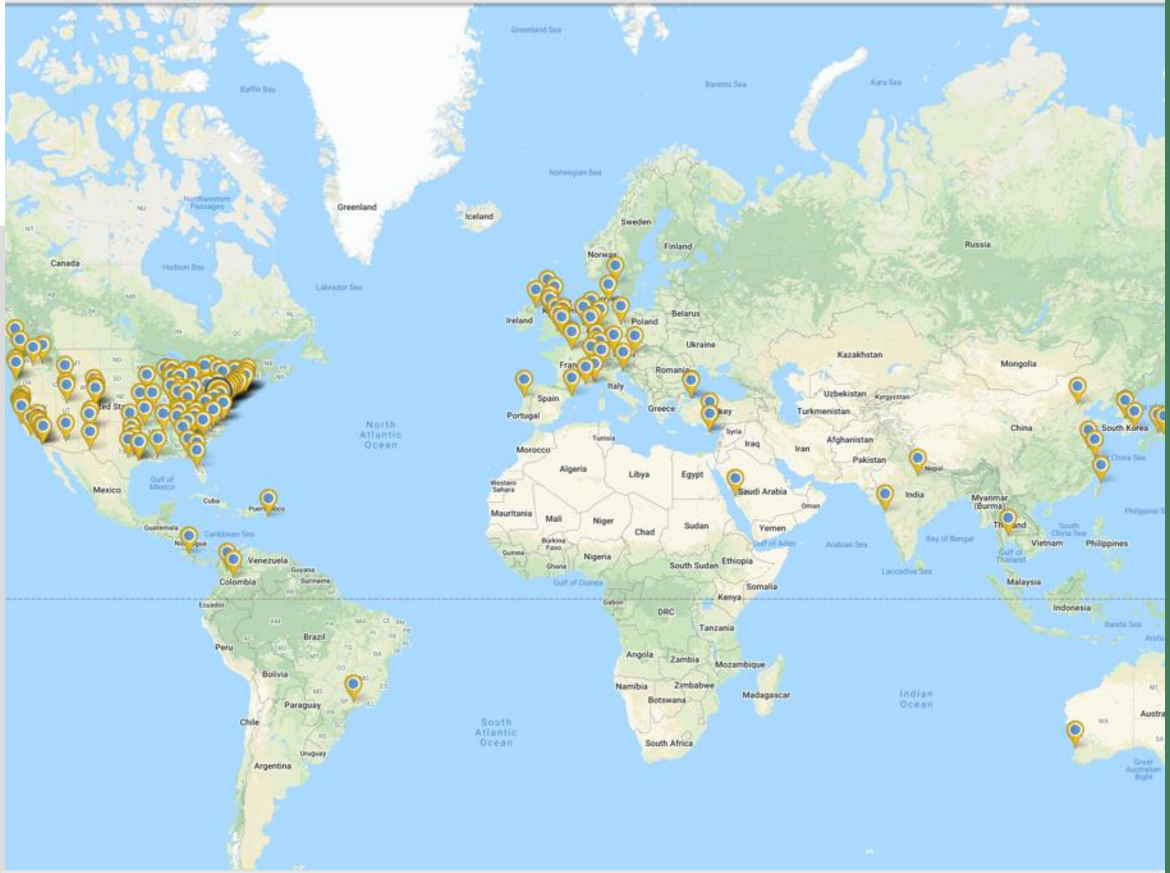
84
patents issued in FY23

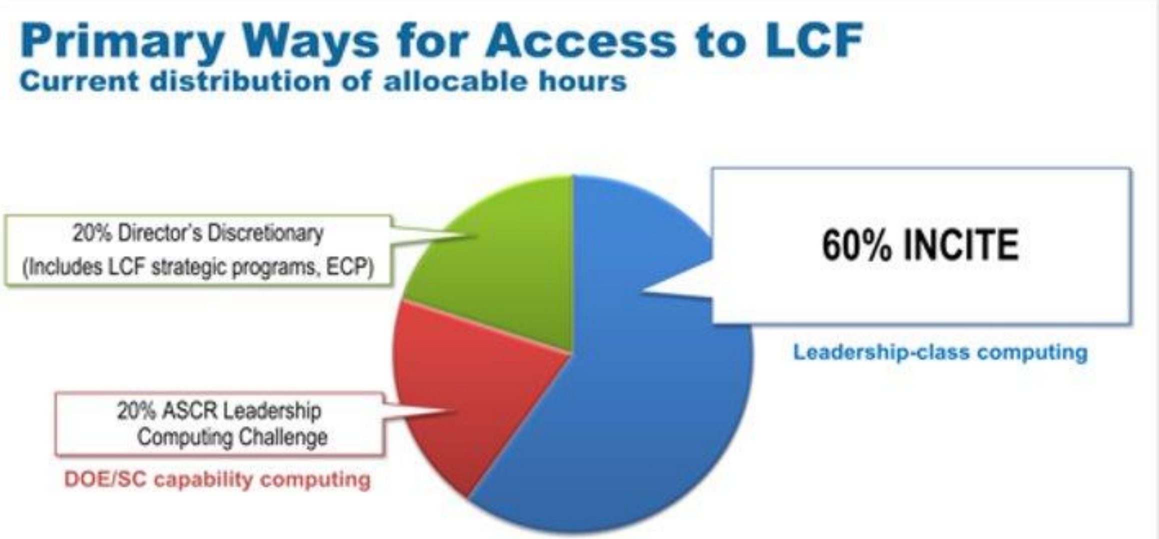
ORNL's mission

Deliver scientific discoveries and technical breakthroughs needed to realize solutions in clean energy and national security and provide economic benefit to the nation

OLCF by the Numbers

- ~1,500 users, located around the world.
- ~250 research projects / year
- OLCF users come from academia, government laboratories, federal agencies, and industry
- OLCF resources are allocated through three highly competitive allocation programs requiring peer reviewed proposals
- Since 2012, the OLCF has enabled ~5,500 publications in open literature
- In 2024, 54% of the cycles on Frontier consumed 20% or more of the total node count



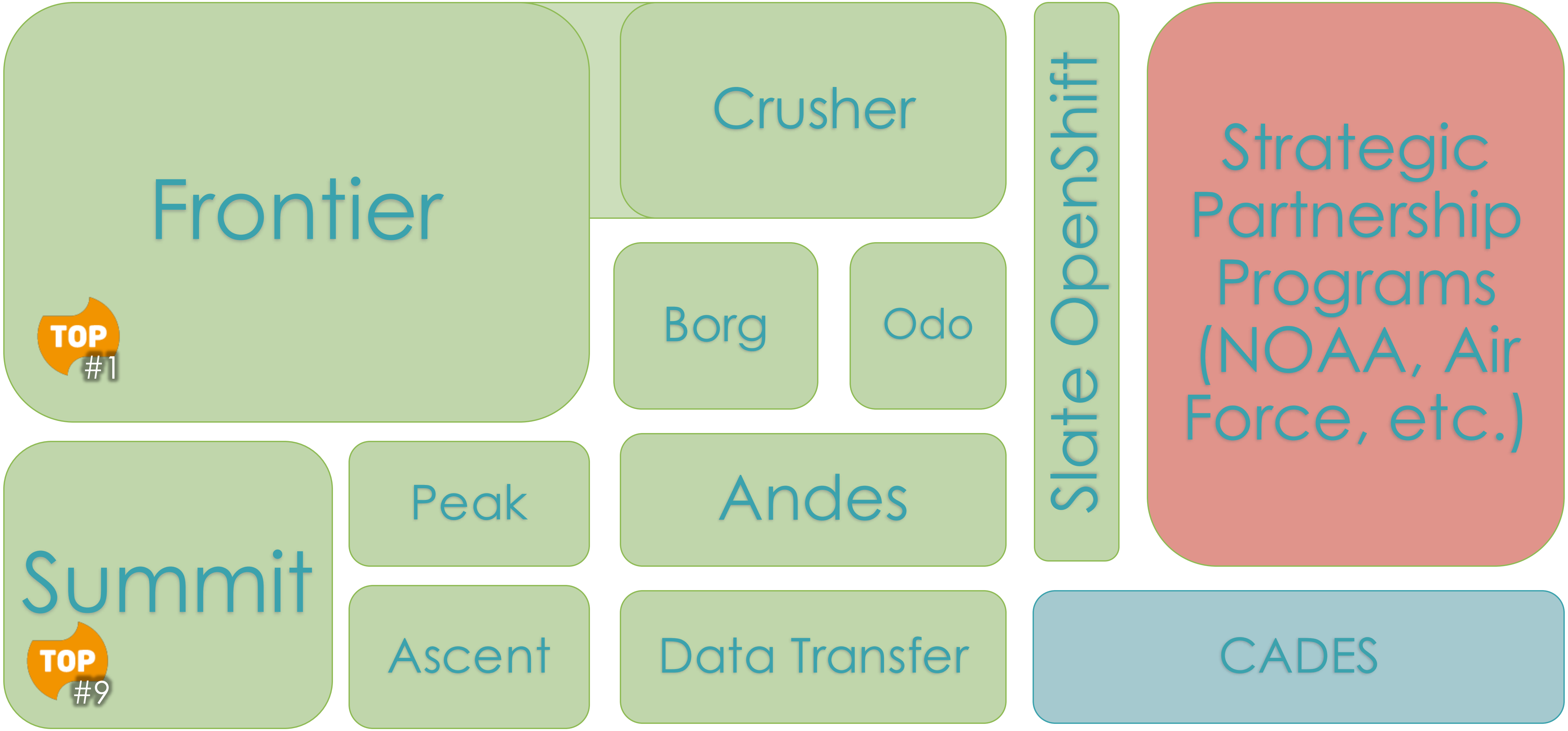


Primary Ways for Access to LCF

Current distribution of allocable hours

20% Director's Discretionary
(Includes LCF strategic programs, ECP)
60% INCITE
Leadership-class computing
20% ASCR Leadership Computing Challenge
DOE/SC capability computing

ORNL NCCS Compute Resources



Frontier

TOP #1

Crusher

Borg

Odo

Slate OpenShift

Strategic Partnership Programs (NOAA, Air Force, etc.)

Summit

TOP #9

Peak

Ascent

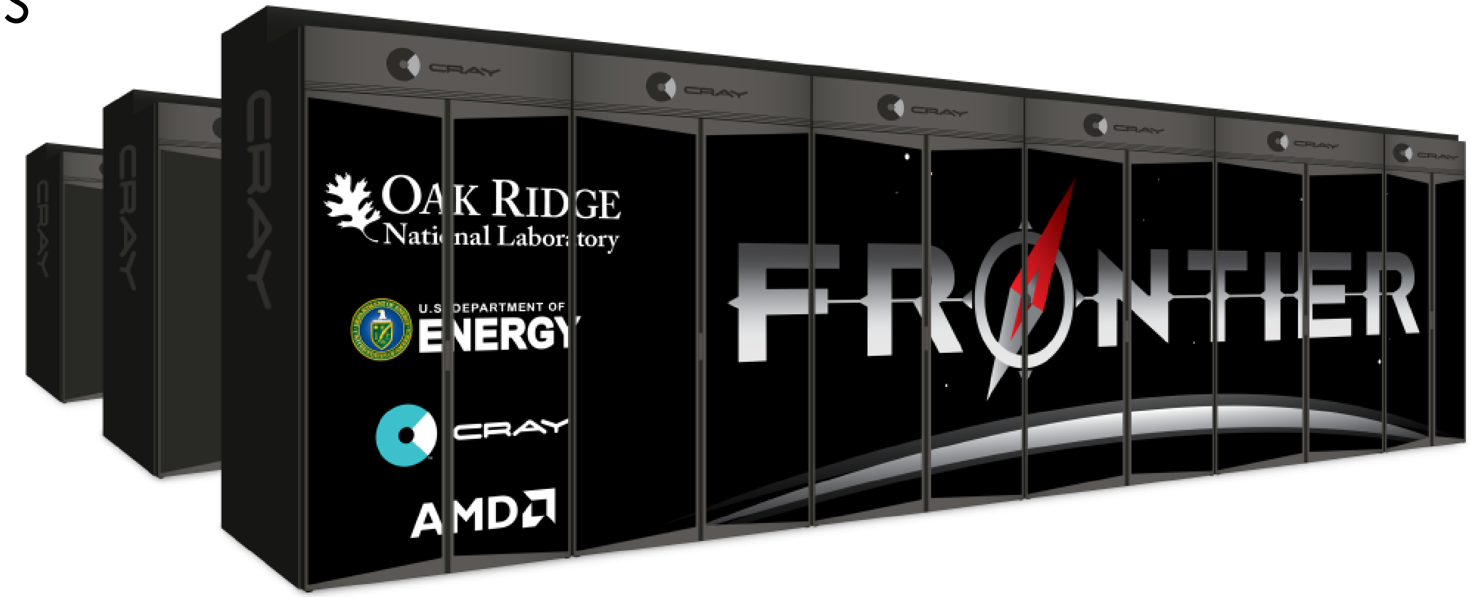
Andes

Data Transfer

CADES

Slurm on Frontier

- Updated to Slurm 24.05 on 8/20/2024
- Exclusive node allocations
- Extensive node health checking scripts run between jobs
  - Nodes that recover are automatically returned
- Slingshot switch plugin
- Custom *jobstat* script show system/job status



Frontier Queue Policy

Bin	Min Nodes	Max Nodes	Max WallTime	Boost (Days)
1	5,645	9,408	12	8
2	1,882	5,644	12	4
3	184	1,881	12	0
4	92	183	6	0
5	1	91	2	0
Extended	1	64	24	0

Debug QOS gives a 2 day boost

Scalable Protected Infrastructure

- Adds the ability to run “moderate data enhanced” workloads on Frontier
- Locks down nodes, adds routes and iptables rules, switches out file systems
- Forces a reboot at the end of the job for cleanup
- Currently implemented as a partition (but looking to move to a “rebootless” node feature)

Slurm 19.05 Sponsored Work

Ticket ID	Title	Notes
4887	Disable setting triggers from non-root/slurm_user by default	Now off by default, add *SlurmctldParameters=allow_user_triggers* to re-enable
5716	Be able to disable resizing of jobs	Now off by default, add *SchedulerParameters=permit_job_expansion* to re-enable
6286	Add priority to associations	
6287	Add ability to set priority factor for job size	Implemented as "site" factor that can be set with a plugin, job_submit, or scontrol
6288	Add ability to not normalize priorities	Added NO_NORM_* flags
	Hand-set priority as a factor, not an override	Can use existing *nice* value (negative nice is a positive boost)

CORAL-2 Slurm 20.02 & 20.11 Sponsored Work

Ticket ID	Title	Notes
7591	Reservation Affinity (magnetic flag)	Slurm 24.11 will reserve in the backfill map with *SchedulerParameters=bf_allow_magnetic_slot* (Ticket 19507)
7561	Provide a REST API to accounting data captured within slurmdbd	Initial implementation of the *dbv#* endpoint
7594	Step-level GPU binding and affinity	
7593	Heterogenous step support	
8573	Interactive step for salloc	
7562	acct_gather_interconnect/sysfs plugin	

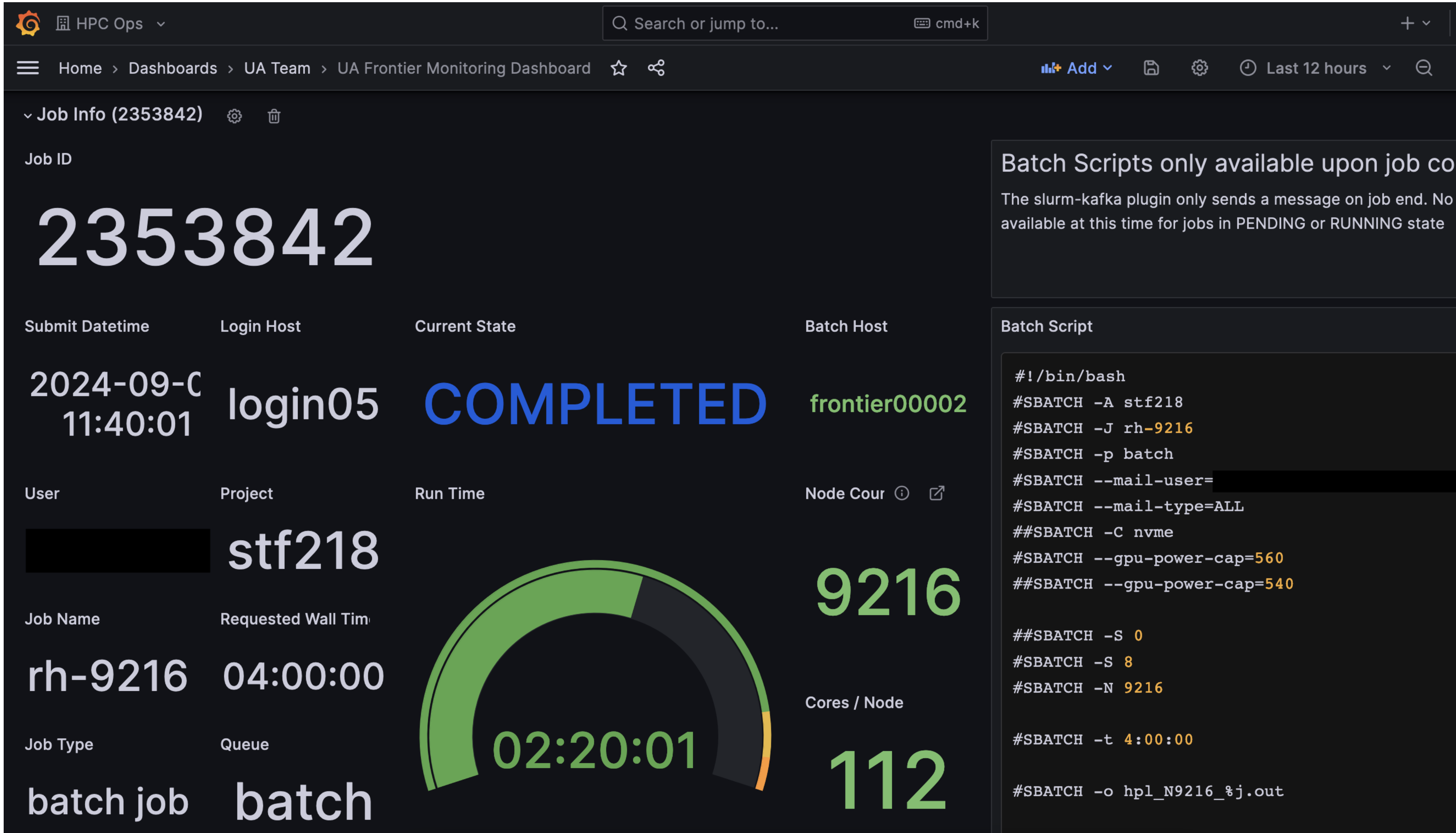
and others…

Slurm 23.02 Sponsored Work

Ticket ID	Title	Notes
13382	User-supplied nodelist larger than requested nodes	
13446	Support reservation nodelist updates with Nodes+= and Nodes-=	
13380	Reservation list currently on a node	Visible with *scontrol show node*
15196	Node failure accounting for jobs	Field is *FailedNode* in *sacct*
10855	Reservation comment field	
15195	Add jobcomp/kafka plugin	Still tracking 19978: jobcomp not including energy

Job Completion - Kafka

- Future work – expand to also send on submission and start? Ticket 20270


HPC Ops
Search or jump to...
cmd+k
Home › Dashboards › UA Team › UA Frontier Monitoring Dashboard
Add
Last 12 hours

Job Info (2353842)

Job ID

2353842

Submit Datetime

2024-09-C
11:40:01

Login Host

login05

Current State

COMPLETED

Batch Host

frontier00002

User

[redacted]

Project

stf218

Run Time

02:20:01

Node Cour

9216

Job Name

rh-9216

Requested Wall Tim

04:00:00

Cores / Node

112

Job Type

batch job

Queue

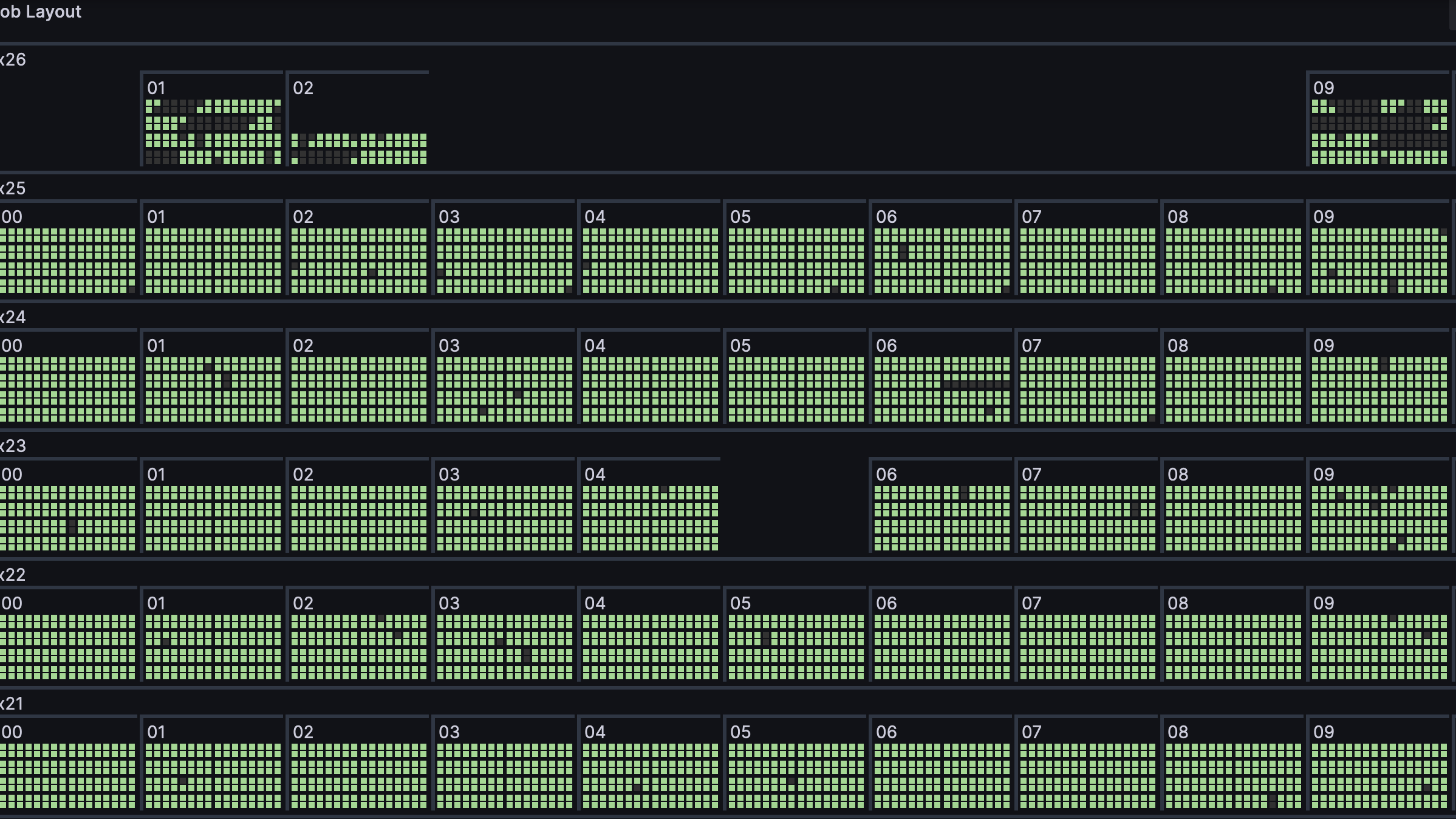
batch

Batch Scripts only available upon job co

The slurm-kafka plugin only sends a message on job end. No
available at this time for jobs in PENDING or RUNNING state

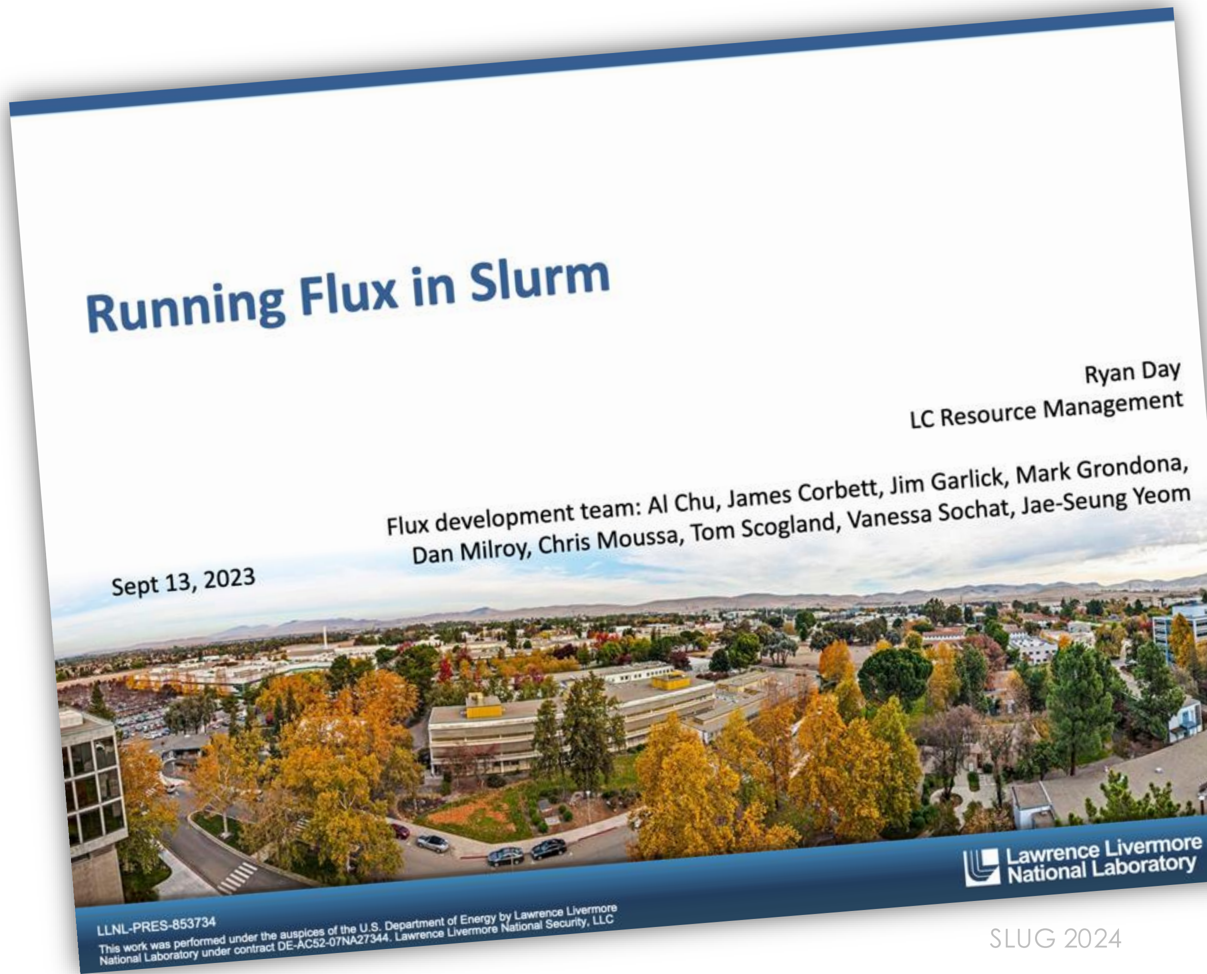
Batch Script

```
#!/bin/bash
#SBATCH -A stf218
#SBATCH -J rh-9216
#SBATCH -p batch
#SBATCH --mail-user=[redacted]
#SBATCH --mail-type=ALL
##SBATCH -C nvme
#SBATCH --gpu-power-cap=560
##SBATCH --gpu-power-cap=540

##SBATCH -S 0
#SBATCH -S 8
#SBATCH -N 9216

#SBATCH -t 4:00:00

#SBATCH -o hpl_N9216_%j.out
```


ob Layout
x26
01
02
09
x25
00
01
02
03
04
05
06
07
08
09
x24
00
01
02
03
04
05
06
07
08
09
x23
00
01
02
03
04
06
07
08
09
x22
00
01
02
03
04
05
06
07
08
09
x21
00
01
02
03
04
05
06
07
08
09

User Parallelism with Many Steps

- Some users express their parallelism with lots of steps instead of using MPI
- Job steps take global locks and excessive step counts can cause slowdown
- We recommend flux to run steps instead



Running Flux in Slurm

Ryan Day
LC Resource Management

Flux development team: Al Chu, James Corbett, Jim Garlick, Mark Grondona, Dan Milroy, Chris Moussa, Tom Scogland, Vanessa Sochat, Jae-Seung Yeom

Sept 13, 2023

Lawrence Livermore National Laboratory

LLNL-PRES-853734
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Isolated Step Manager

- Brian presented about the step manager improvements last year
- Ran into some issues with the Slingshot plugin and mpi/cray_shasta – resolved before 24.05
- Still waiting on scale tests on Frontier

Step Management Enhancements

Brian Christiansen

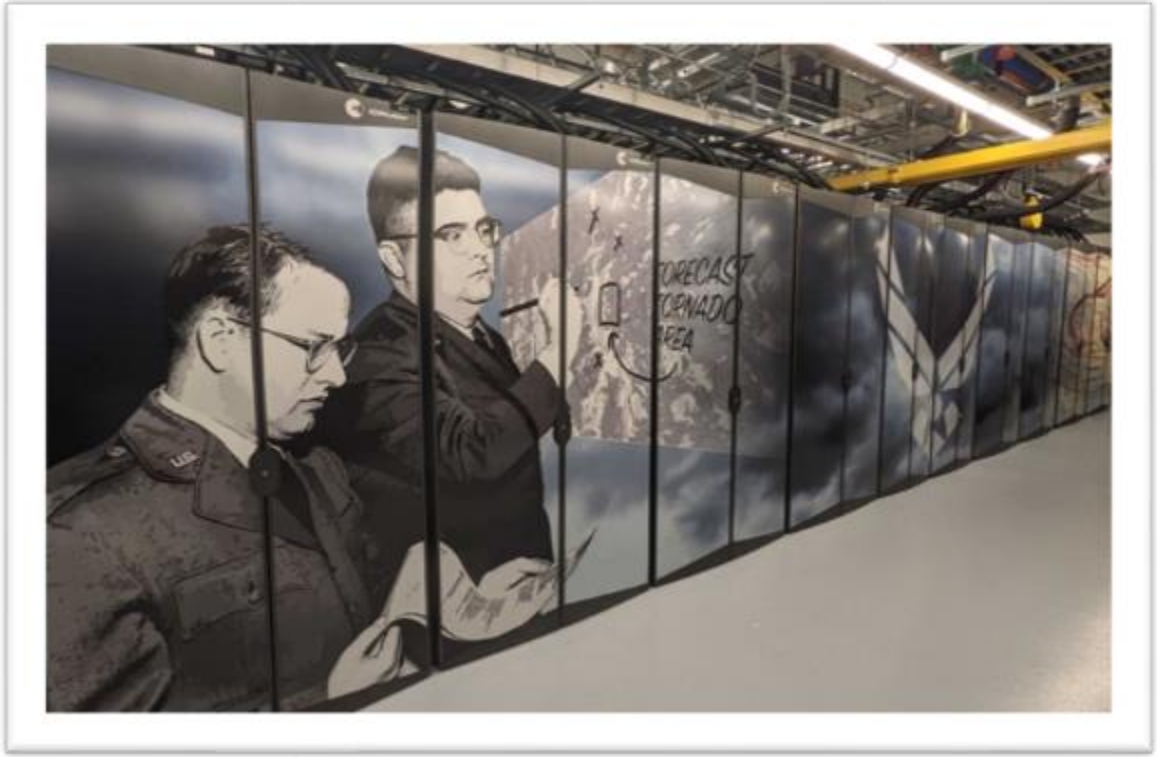
Slurm User Group 2023

AFW HPC 11

UNITED STATES AIR FORCE
WEATHER OPERATIONS
FORECAST
TORNADO
AREA

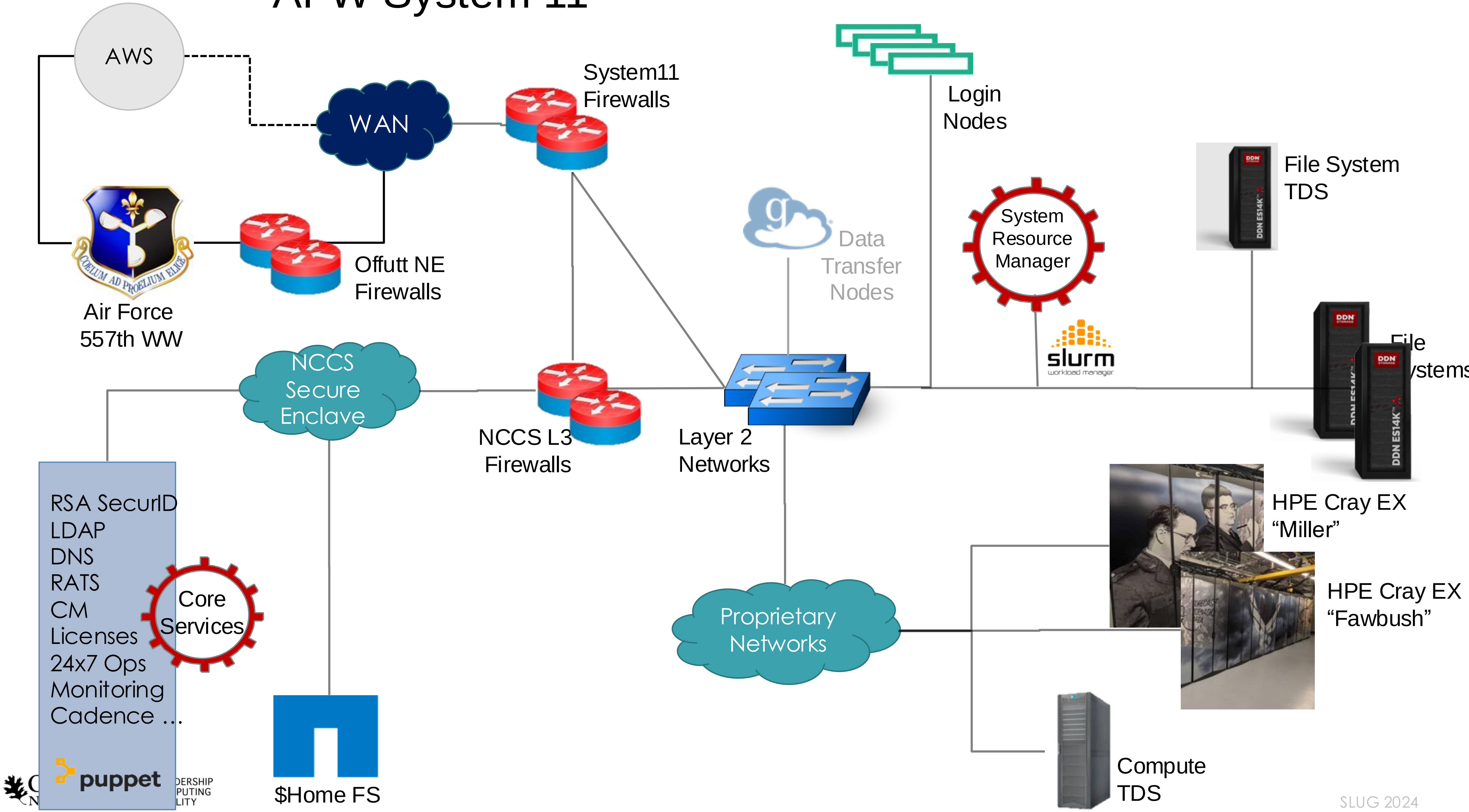
System 11 Computing Capability

- Provides 6.5x the sustained computing capacity of existing AFW HPCs capabiity
- Redundant compute halls & file systems provide highly available and flexible services
- Federated workload manager seamlessly manages the multiple compute partitions without user intervention.
- Both compute & storage are easily scalable beyond the baseline configuration
- The addition of NVIDIA GPUs supports develop/test of next generation models & algorithms and AFW AI/ML efforts



AFW's System 11 – Redundant 800-node Cray EX supercomputers managed by ORNL provide more than 6x the capability of their previous system. Initial operational capability in 2021.

AFW System 11



AWS

WAN

System11 Firewalls

Login Nodes

File System TDS

Air Force 557th WW

Offutt NE Firewalls

Data Transfer Nodes

System Resource Manager

slurm workload manager

File ystems

NCCS Secure Enclave

NCCS L3 Firewalls

Layer 2 Networks

RSA SecurID
LDAP
DNS
RATS
CM
Licenses
24x7 Ops
Monitoring
Cadence …

Core Services

puppet

HPE Cray EX "Miller"

HPE Cray EX "Fawbush"

Proprietary Networks

$Home FS

Compute TDS

Slurm on AFW HPC11

High Capacity

- 2x Hall Design
- Double Compute Capacity

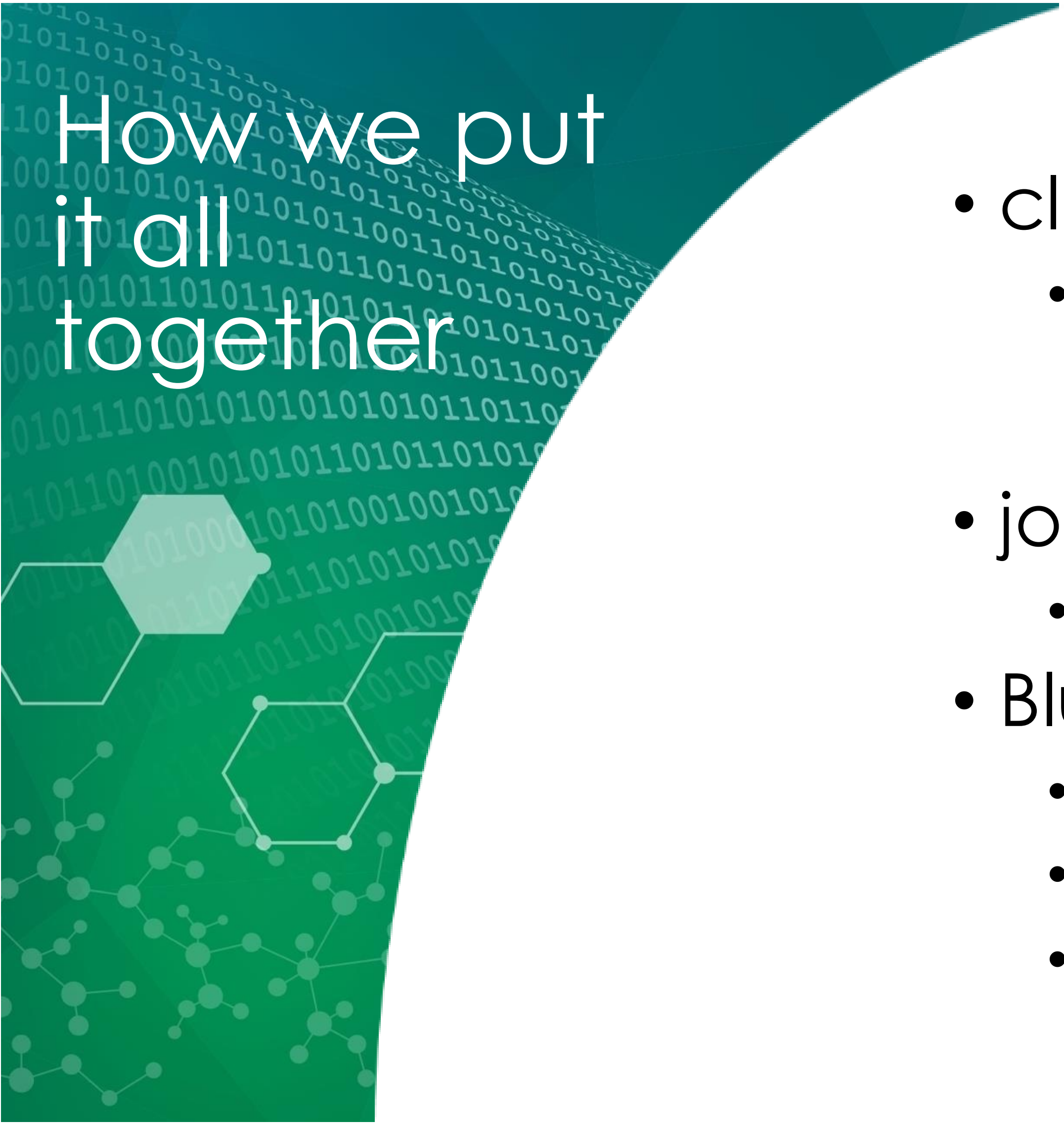
High Availability

- Jobs co-scheduled on both systems
- Allows ORNL to take down individual halls without disrupting operations

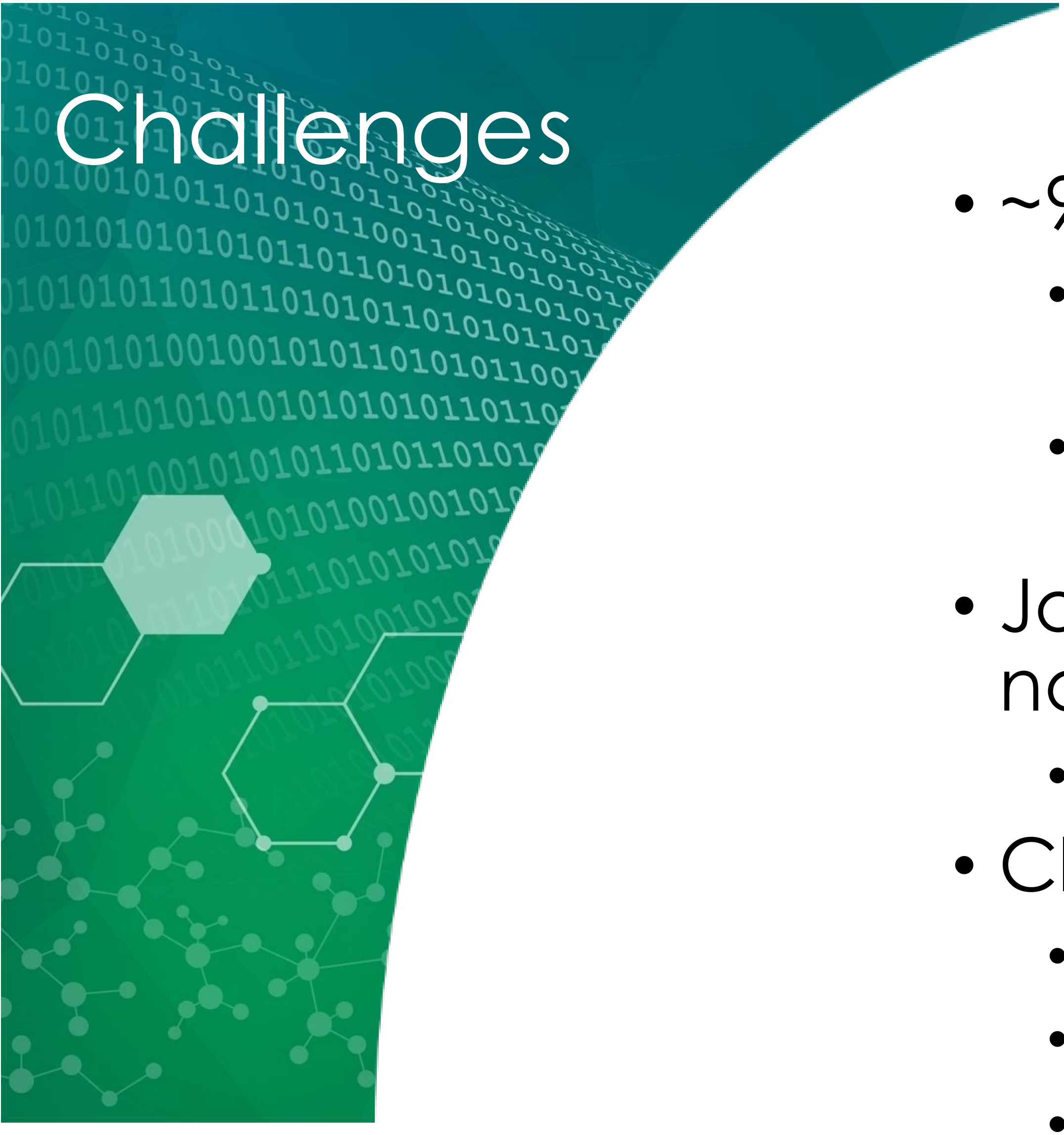
Scrontab

- Most production workflows are scheduled through scron
- Weather forecasting is done at explicit time intervals

How we put it all together



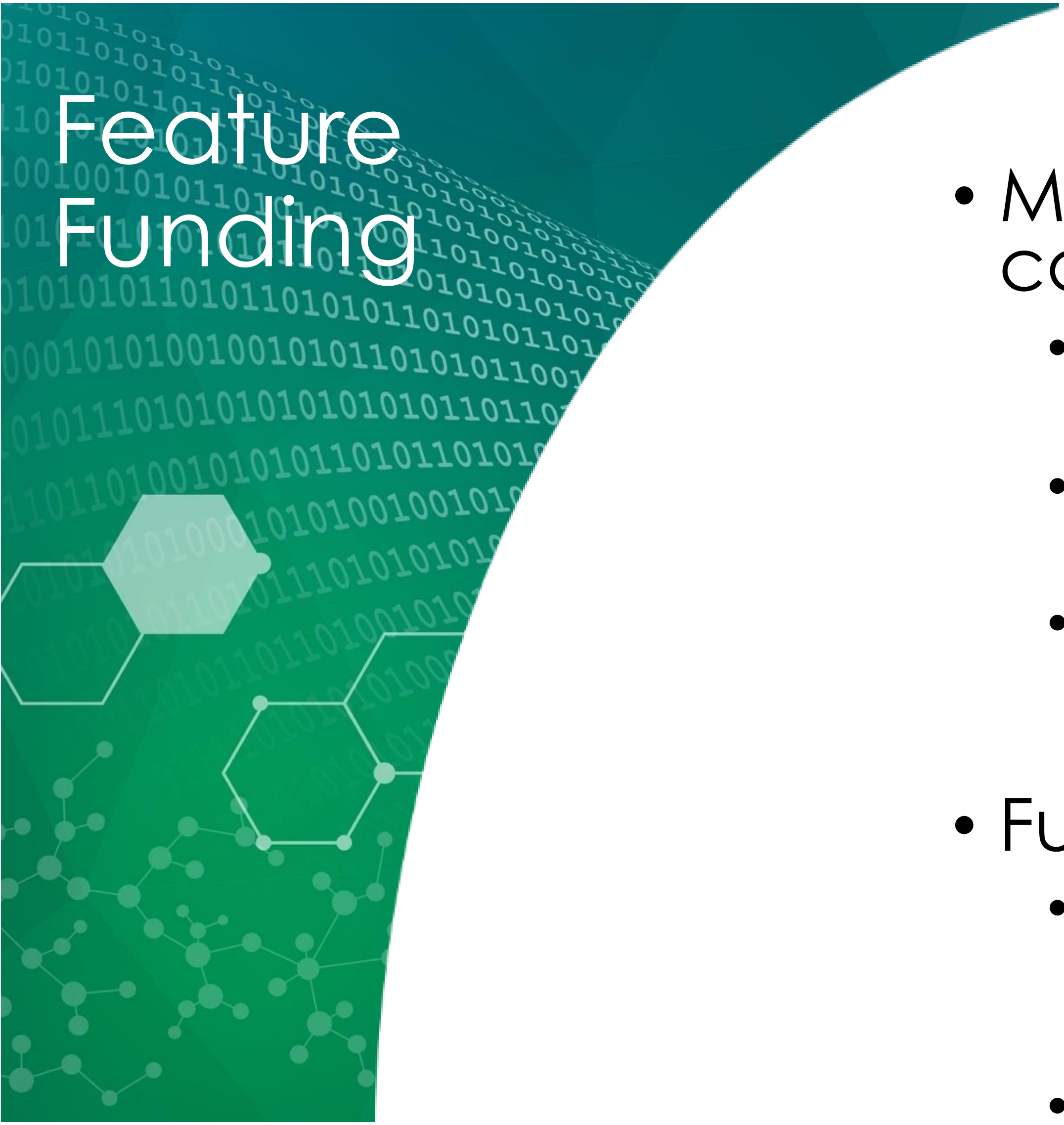
- cli_filter
  - Ensure jobs will actually run, enforce timelimits, set clusters, enforce cluster-constraints
- job_submit
  - Belt and suspenders for specific items
- Blue/Green deployments
  - Cluster constraint
  - Login cluster as a node constraint
  - System upgrades and default programming environment changes



Challenges

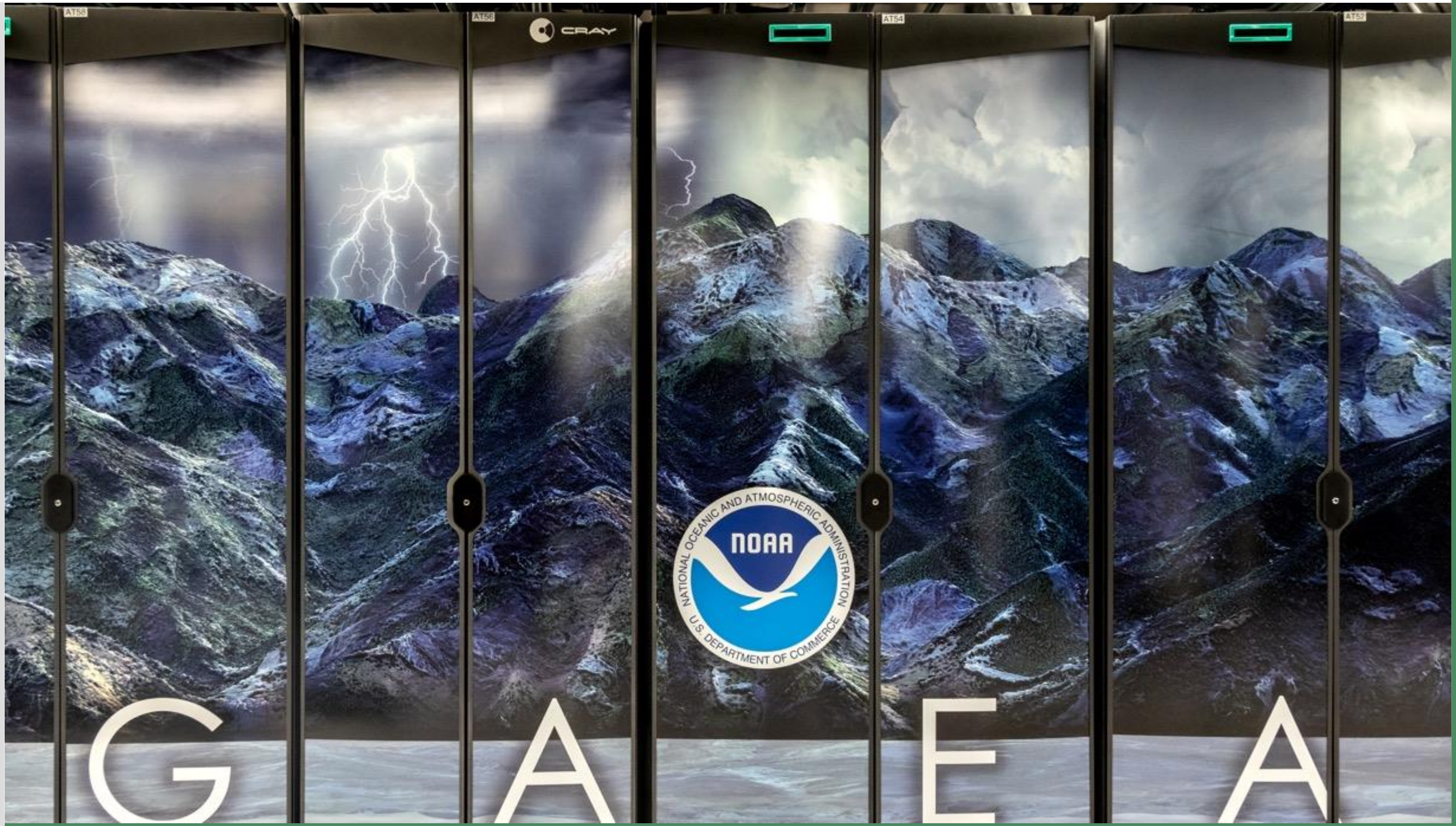
- ~90K jobs submitted per day
  - +95% of those are less than one node jobs
  - Node sharing is enabled at a per account level
- Jobs are scheduled by time interval not queue depth
  - Sawtooth cluster utilization
- Cluster utilization imbalance
  - Primary sibling
  - Scheduler cycle runs faster
  - First come

Feature Funding



- Make federated slurm more HA capable
  - Remove the single point of failure of the slurmdbd
  - Prior to 24.05 all client communications were routed through the slurmdbd
  - Allows for the slurmdbd to be upgraded without breaking all job submissions and client command use
- Future work
  - Allow for individual clusters to be updated without having to update both at the same time
  - Federated job scheduling has issues when controllers are running two different versions

NOAA


CRAY
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE
G A E A

National Climate-Computing Research Center (NCRC)

- Agreement between NOAA and DOE's Oak Ridge National Laboratory for HPC services and climate modeling support
- Strategic Partnership Project, currently in year 13
- 5-year periods. Current IAA effective through FY25
- Within ORNL's National Center for Computational Sciences (NCCS)
- Service provided - DOE-titled equipment
- Secure network enclave; Department of Commerce access policies
- Gaea
  - 2x HPE EX (~2K Compute Nodes), 2x GPFS (65PB)
  - Mission: R&D, long-term climate and weather predictions and projections

Federated but not federated

- cli_filter
  - Ensure jobs will actually run, enforce timelimits, set clusters
- job_submit
  - Belt and suspenders for specific items
- Federated View, but not federated jobs
  - Multi-Cluster
  - C5 and C6 have different projects and missions associated with them
  - Slightly different processor generations

Feature Funding

- Stdout through sacct

```
gaea61:~ # sacct -j 135144112 --format=stdin,stderr,stdout
               StdIn               StdErr               StdOut
-------------------- -------------------- --------------------
           /dev/null                      /gpfs/f5/gfdl_m/scr+
```

Future Slurm Improvement Ideas

- Support subuid/subgid ranges in nss_slurm (Ticket 19551)
- Improvements to `sbcast --send-libs` to reduce stats for excluded libraries (Ticket 20270)
- RPM spec file to support patches and custom version “release” numbers (Ticket 20555)
- Integrate JupyterHub using slurmrestd instead of SSH or batch spawner

IRI Blueprint Science Patterns

Time-Sensitive Pattern

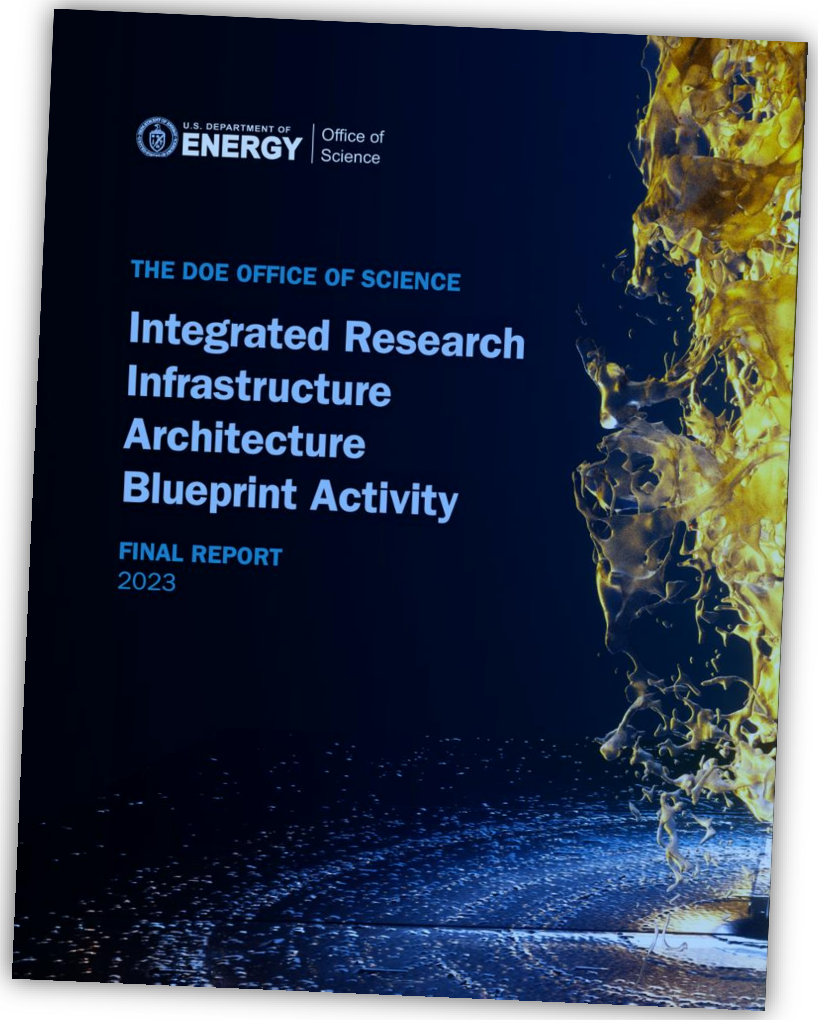
- Workflows that have time-critical requirements (i.e., real time) motivated by factors including rapid decision-making, experiment control, coordinating distributed assets, and data capture/reduction

Data Integration-Intensive Pattern

- Analysis of data from multiple sources, e.g., simulations and experiments/observations
- Cross-site data-driven discovery
- AI/ML incorporated into simulations and experiments

Long-Term Campaign Pattern

- Sustained access (several years) to resources at scale, e.g., sustained simulation production and large data (re)processing for collaborative use



U.S. DEPARTMENT OF ENERGY | Office of Science

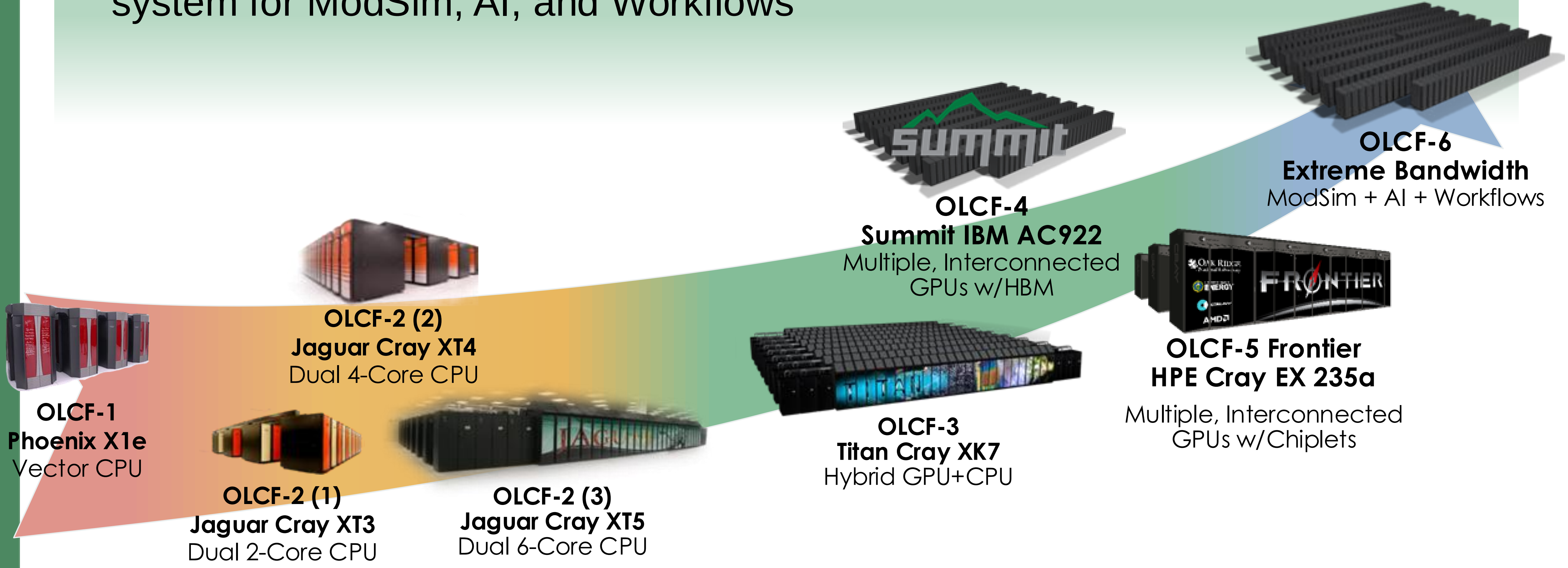
THE DOE OFFICE OF SCIENCE

Integrated Research Infrastructure Architecture Blueprint Activity

FINAL REPORT
2023

OLCF Drives Technology Innovation at Scale

- OLCF has pushed the boundaries on performance at scale
- OLCF-6 will push the boundaries on bandwidth throughout the system for ModSim, AI, and Workflows



OLCF-1
Phoenix X1e
Vector CPU

OLCF-2 (1)
Jaguar Cray XT3
Dual 2-Core CPU

OLCF-2 (2)
Jaguar Cray XT4
Dual 4-Core CPU

OLCF-2 (3)
Jaguar Cray XT5
Dual 6-Core CPU

OLCF-3
Titan Cray XK7
Hybrid GPU+CPU

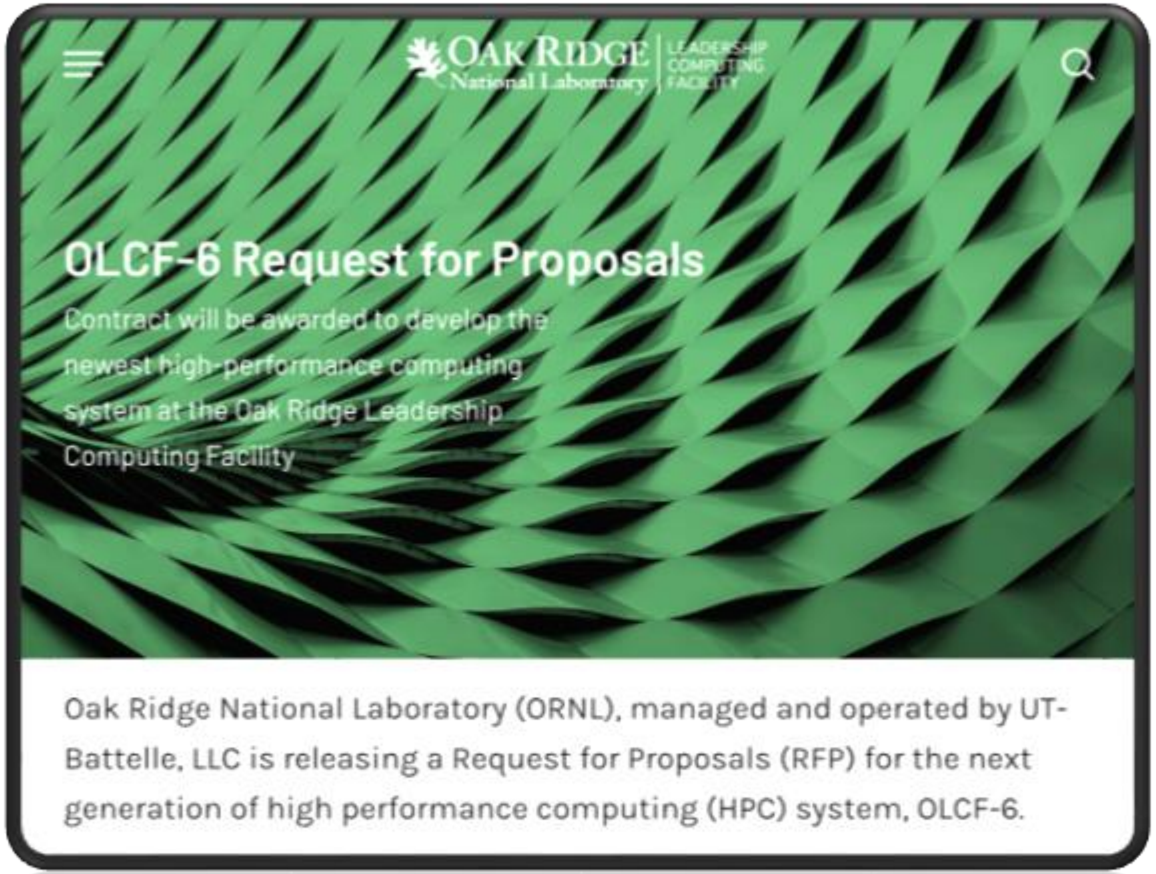
OLCF-4
Summit IBM AC922
Multiple, Interconnected GPUs w/HBM

OLCF-5 Frontier
HPE Cray EX 235a
Multiple, Interconnected GPUs w/Chiplets

OLCF-6
Extreme Bandwidth
ModSim + AI + Workflows

Up Next at ORNL - Discovery

- Proposals were due August 30, 2024
- Reviews are in progress



OLCF-6 Request for Proposals

Contract will be awarded to develop the newest high-performance computing system at the Oak Ridge Leadership Computing Facility

Oak Ridge National Laboratory (ORNL), managed and operated by UT-Battelle, LLC is releasing a Request for Proposals (RFP) for the next generation of high performance computing (HPC) system, OLCF-6.

Up Next at ORNL - Discovery

6.2.1 Workload Management Features

Offeror will provide a full-featured workload manager with native step management. Company currently uses Slurm and expects that the proposed solution will provide the features available in Slurm 23.11 (or later). If an older version of Slurm or an alternative workload manager is proposed, the Offeror will provide a detailed analysis of the differences between the Offeror's proposed solution and Slurm.

Offeror will ensure that the proposed workload manager supports the full functionality of the proposed system design, including process affinity, accelerator support, and high-speed network features.

Priority: TR-1

Questions?

ezellma@ornl.gov
peltzpl@ornl.gov