

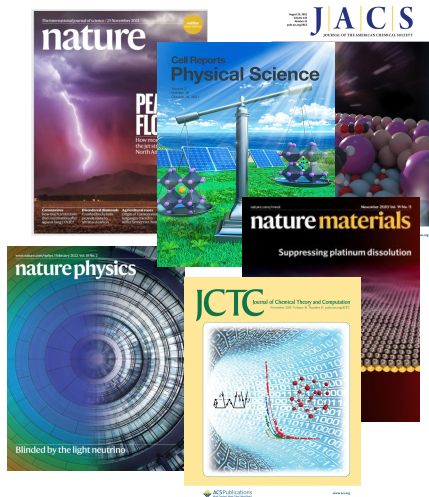
# A View of the Edge from NERSC

April 27th, 2023

Katie Antypas with heavily borrowed  
content and contributions from Debbie  
Bard, Taylor Groves, Ron Kumar, Hai  
Ah Nam, Jay Srinivasan, Rollin Thomas  
and Nick Wright

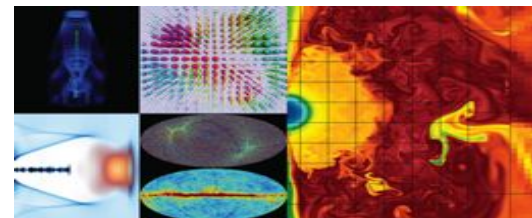
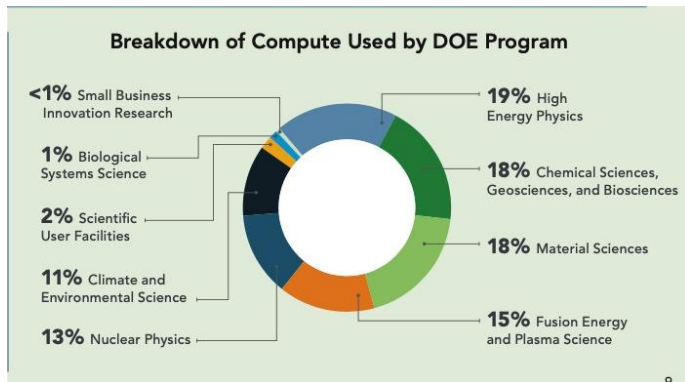
# NERSC is the mission High Performance Computing facility for the DOE Office of Science

9,000 Users  
1,000 Projects

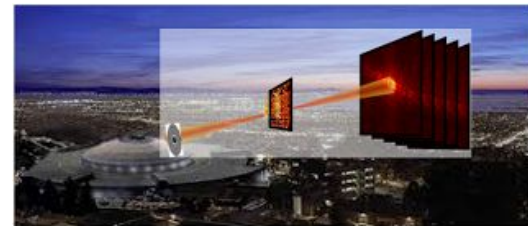


>2,000

Scientific Journal  
Articles per Year

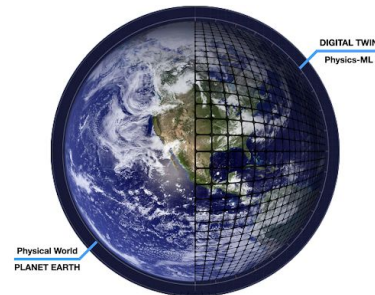


Simulations at scale



Urgent and interactive computing

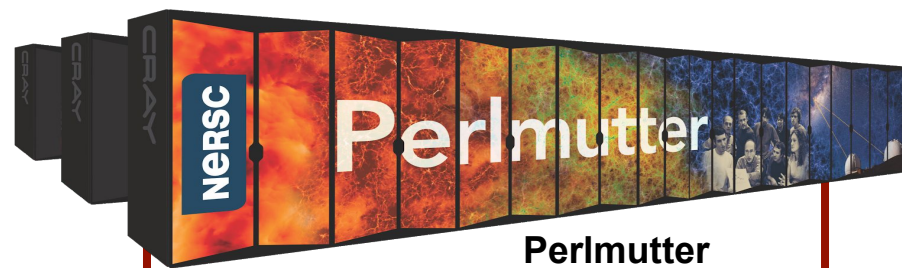
Photo Credit: CAMERA



Complex experimental & AI workflows

Photo credit: A depiction of digital twin Earth adapted from the EU's Destination Earth project.

# NERSC Center Architecture



**Perlmutter**

5 TB/s

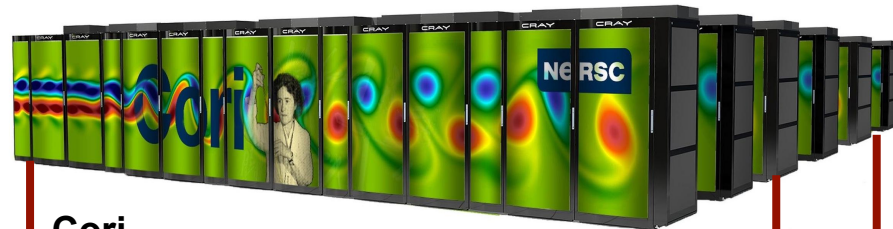
35 PB All-Flash Scratch

**1,792 GPU-accelerated nodes**  
 4 NVIDIA A100 GPUs + 1 AMD "Milan" CPU  
 488 TB (CPU) + 280 TB (GPU) memory

**3,072 CPU-only nodes**  
 2 AMD "Milan" CPUs  
 1,536 TB CPU memory

**HPE Slingshot 11 ethernet-compatible interconnect**  
 4 NICs/GPU node, 1 NIC/CPU node

**TOP 500**  
 The List.  
**#7, 93.8PF Peak**



**Cori**

9,600 Intel Xeon Phi "KNL" manycore nodes  
 2,000 Intel Xeon "Haswell" nodes  
 700,000 processor cores, 1.2 PB memory  
 Cray XC40 / Aries Dragonfly interconnect  
 30 PF Peak

1.5 TB/s

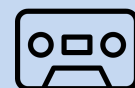
700 GB/s

2 PB Burst Buffer

28 PB Scratch

50 GB/s

**HPSS Tape Archive**  
 ~200 PB



DTNs, Spin, Gateways

**Ethernet & IB Fabric**

Science Friendly Security  
 Production Monitoring  
 Power Efficiency

**LAN**

**ESnet**  
 ENERGY SCIENCES NETWORK

2 x 400 Gb/s  
 2 x 100 Gb/s

100 GB/s

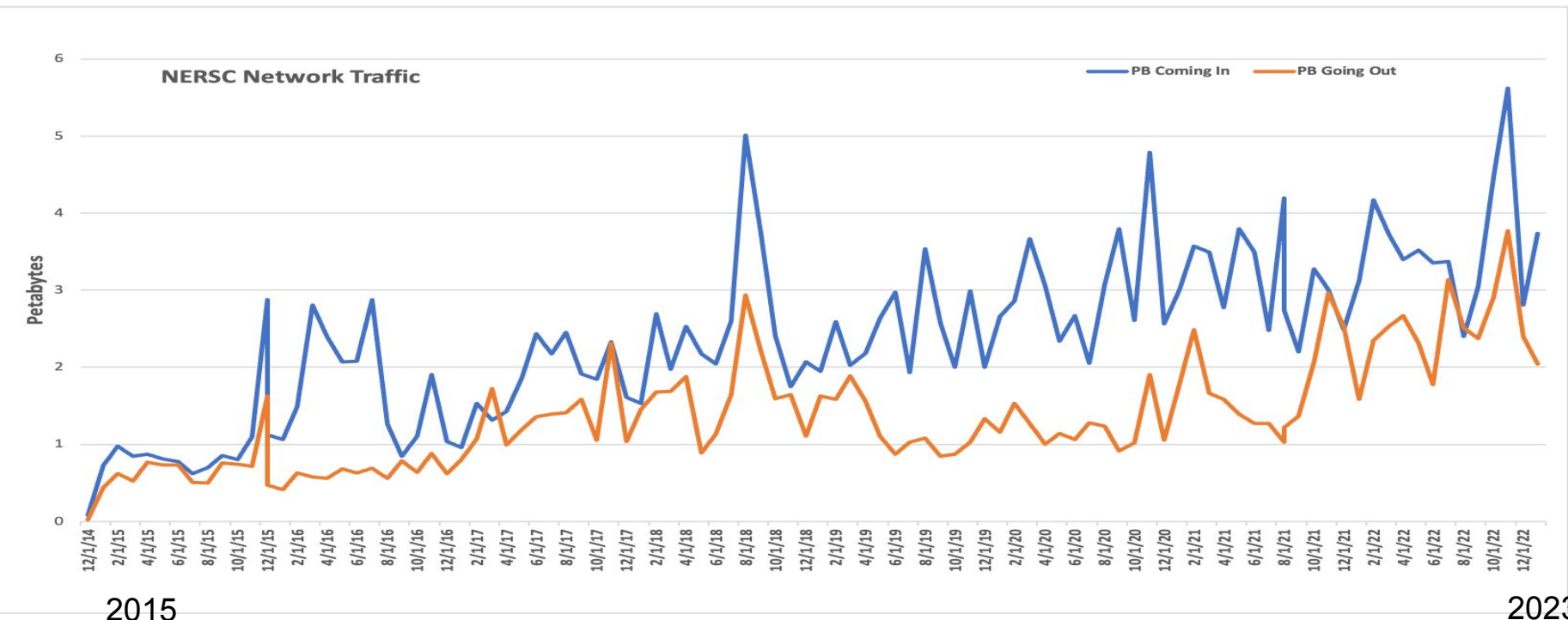
120 PB  
 Community  
 File System

5 GB/s

275 TB  
 /home



# NERSC Border Traffic Per Month

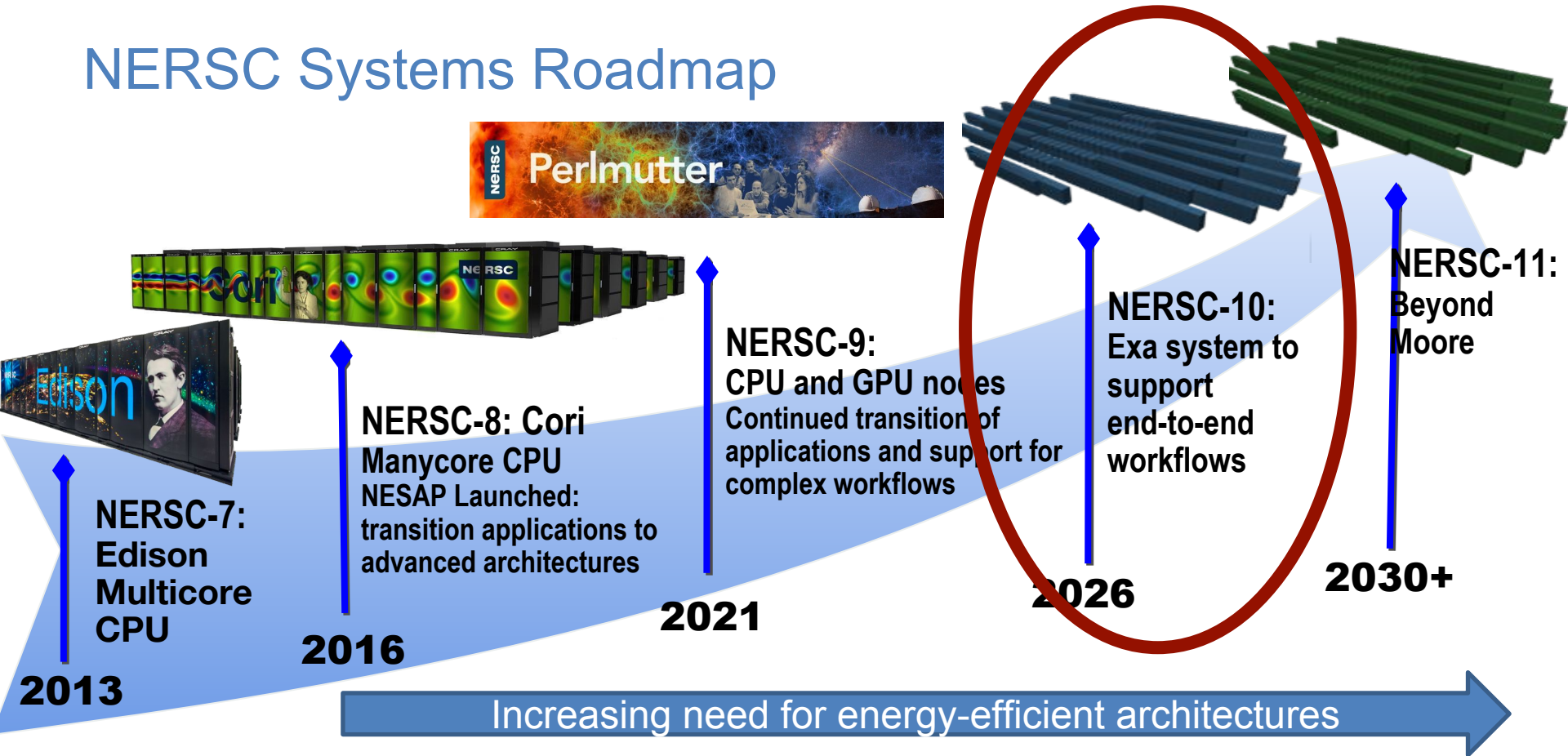


2015

2023



# NERSC Systems Roadmap

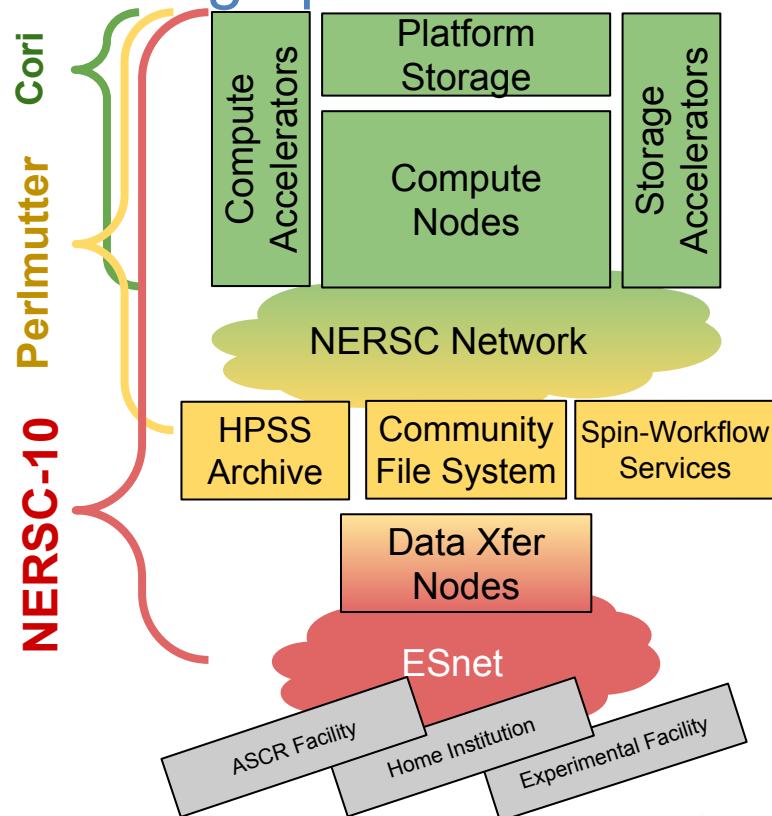


# NERSC-10 Architecture: Designed to support complex simulation and data analysis workflows at high performance

***NERSC-10 will provide on-demand, dynamically composable, and resilient workflows across heterogeneous elements within NERSC and extending to the edge of experimental facilities and other user endpoints***

*New focus in tech specs*

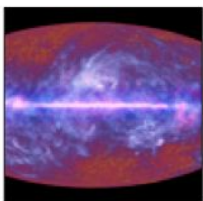
- dynamic orchestration*
- containerization*
- end-to-end workflow performance*
- quality of service*



# NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities



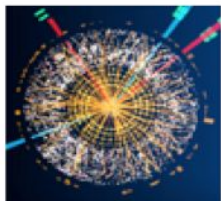
Palomar Transient Factory Supernova



Planck Satellite Cosmic Microwave Background Radiation



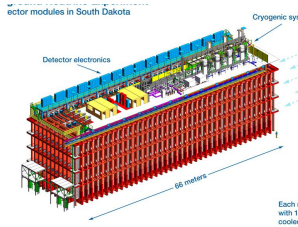
Star Particle Physics



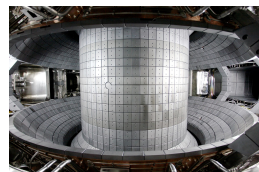
Atlas Large Hadron Collider



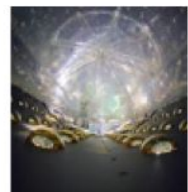
APS



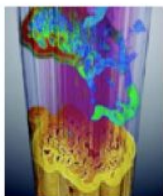
DUNE



KStar



Dayabay Neutrinos



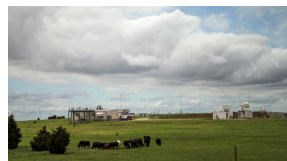
ALS Light Source



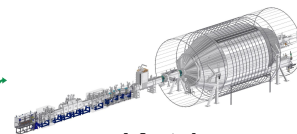
LCLS Light Source



Joint Genome Institute Bioinformatics



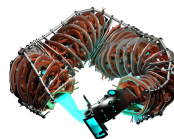
ARM



Katrin



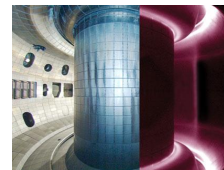
NSLS-II



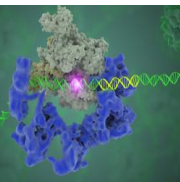
HSX



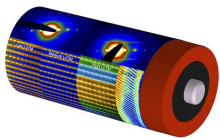
Majorana



DIII-D



Cryo-EM



NCEM

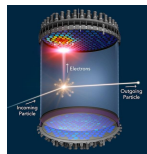


DESI



LSST-DESC

7



LZ



IceCube

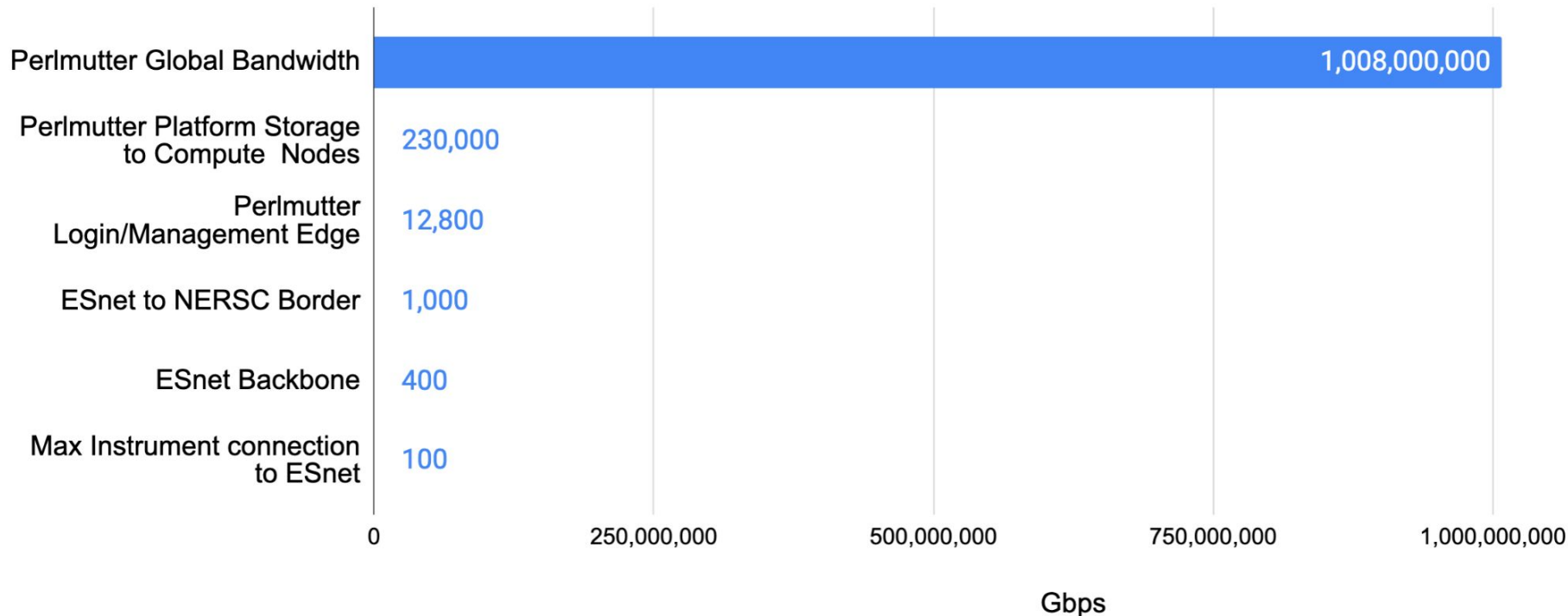


EXO

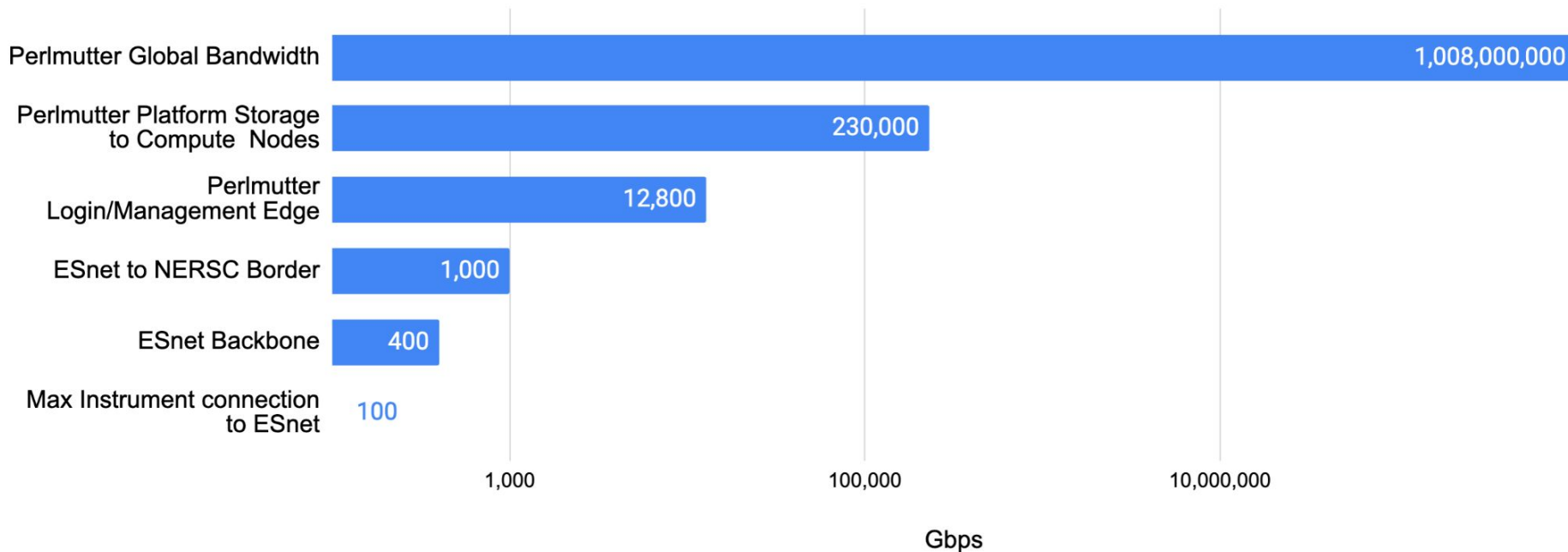




# A View of the Edge from NERSC

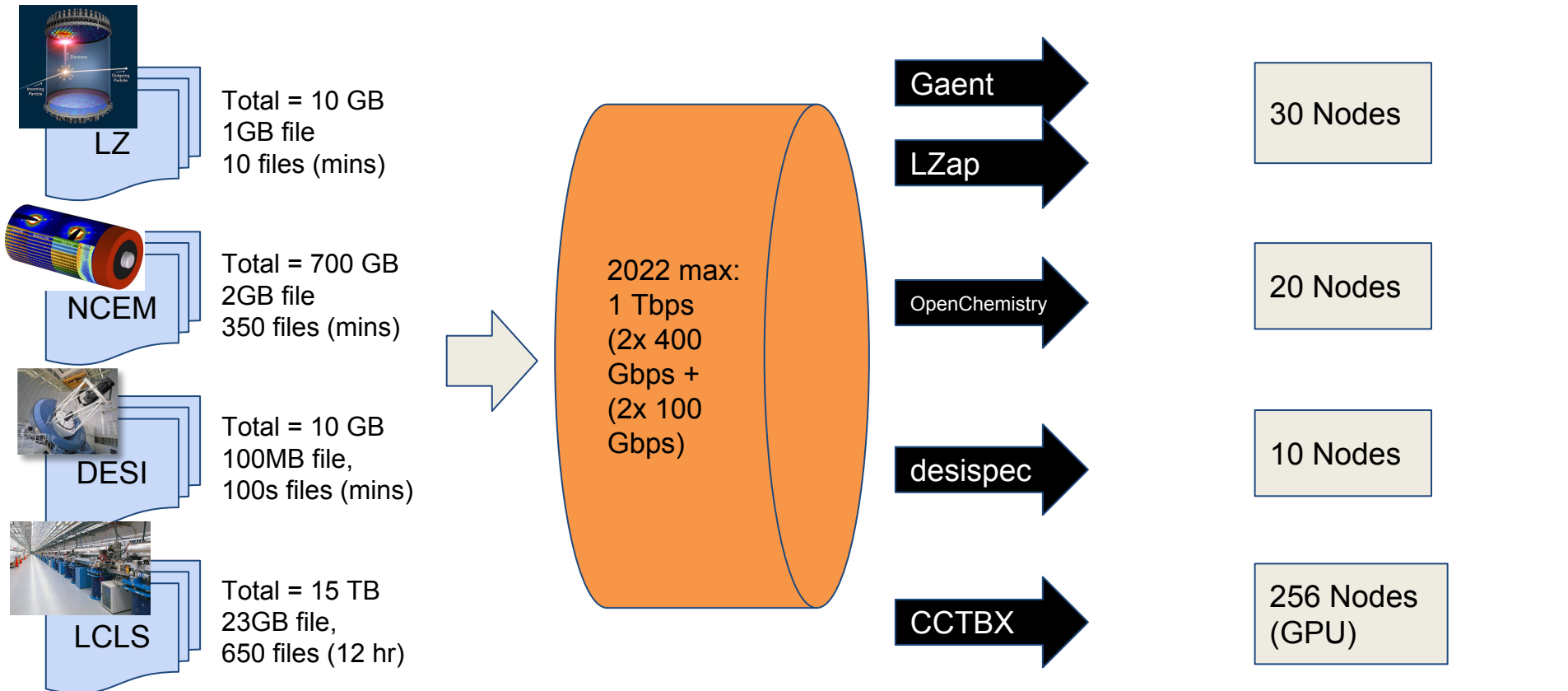


# A Log View of the Edge from NERSC



# Current realtime workload mapping

Credit Debbie Bard and Hai Ah Nam

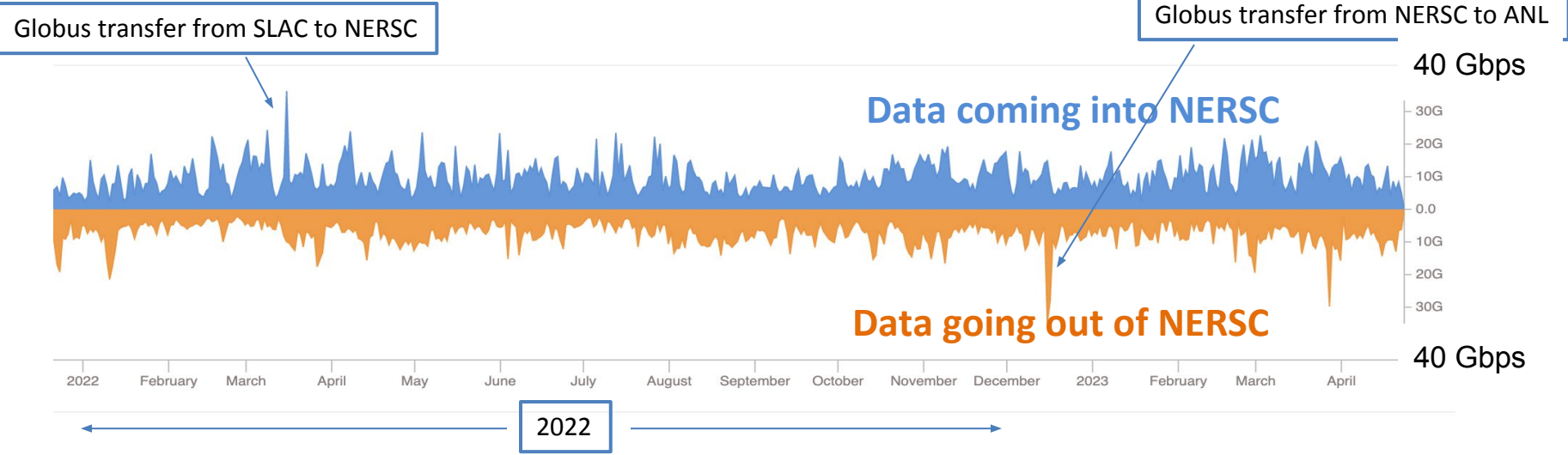


~10GB/sec peak 80Gb/sec peak  
*Easily fits within current network capabilities*

~320 nodes  
*Easily fits within current realtime+reservation capabilities*

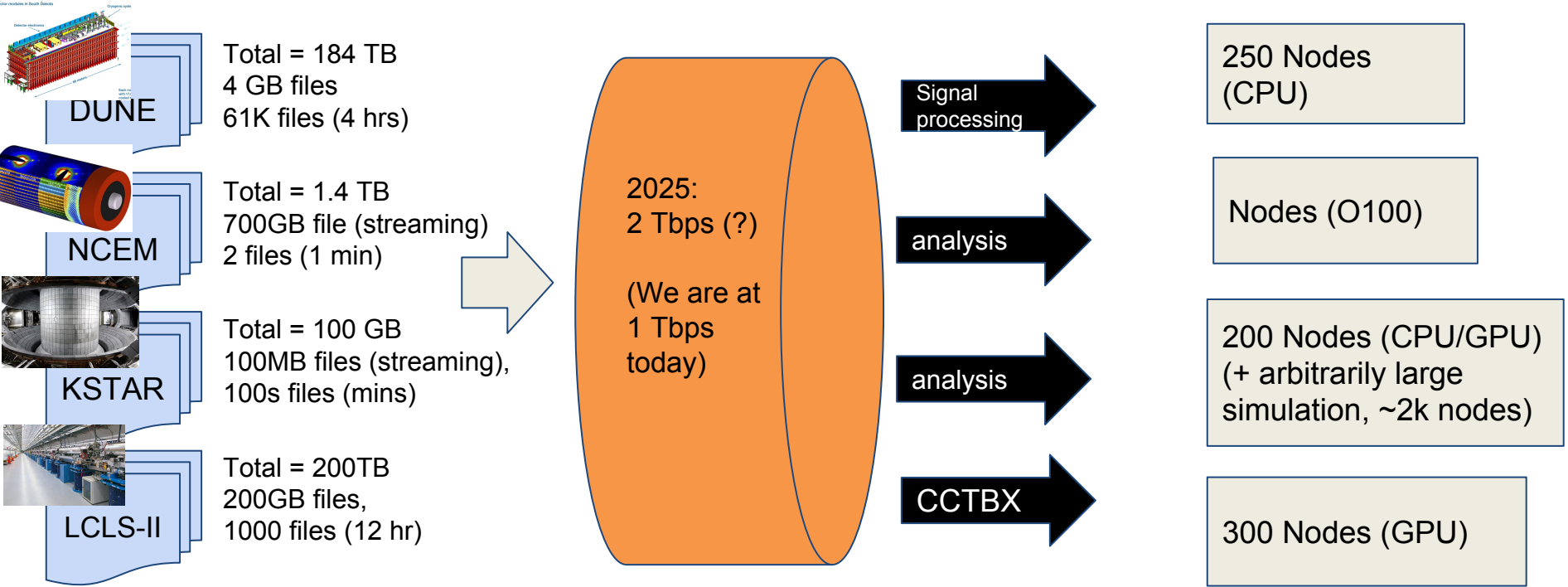


# NERSC Border Traffic



# 2025 realtime workload mapping (rough estimates for subset of top expected realtime workloads)

Credit Debbie Bard and Hai Ah Nam



~45 GB/sec peak 360 Gbps peak [~2.6 TB in mins]  
12

~1000 nodes  
*Will need more if we can shovel the data in fast enough*

# Intermediate Conclusions

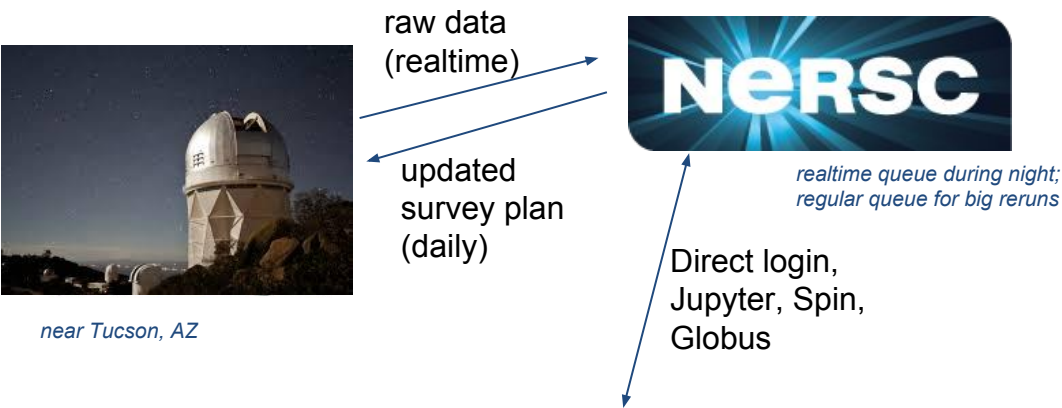
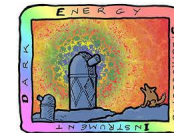
- Experiments are rate-limited based on their own edge bandwidth to ESnet and are already filtering what they send to NERSC. Typical connections 100 Gbps, some even lower. They'd like to transfer data faster if possible.
- If experiments can get data to NERSC, then we can process it
- Yes, there is some nuance - an experiment could send a small input file and need to rerun a simulation, but this use case appears to be less turn-around time sensitive
- We are tracking new experiments coming online and engaging with communities early to understand their requirements.



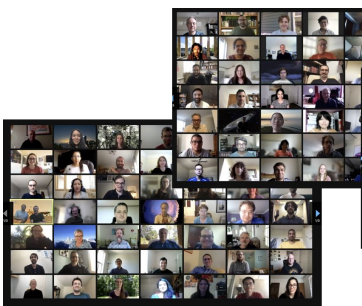
# So then ... why do these collaborations use NERSC?

- Augment local computing and data capability
- Because they have no/little experiment compute/data capabilities
- Store big community data sets for offline analysis and simulations
- For the ecosystem of tools and capabilities: their workflows use a whole suite of NERSC services, computing, data storage, spin for hosted gateways, Jupyter, which would be too difficult/costly to duplicate locally

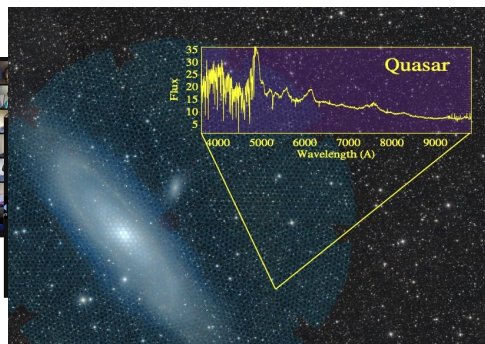
# DESI uses NERSC for nightly data processing



near Tucson, AZ



hundreds of collaborators, worldwide



- NESAP for code optimization:
  - 2.5x improvement in per-node throughput using Perlmutter A100 compared to Cori V100 GPU (x25 compared to Edison).
- Realtime/advanced scheduling for nightly data processing
  - need to process up to 100 GB/night before breakfast to guide telescope operations
- Spin used to monitor data quality and analysis

**Biggest remaining challenge:**  
Robustness / Resilience, especially “soft” outages, e.g. transient I/O or slurm failures  
*Maximizing science is different than maximizing FLOPS or CPU-hours delivered*

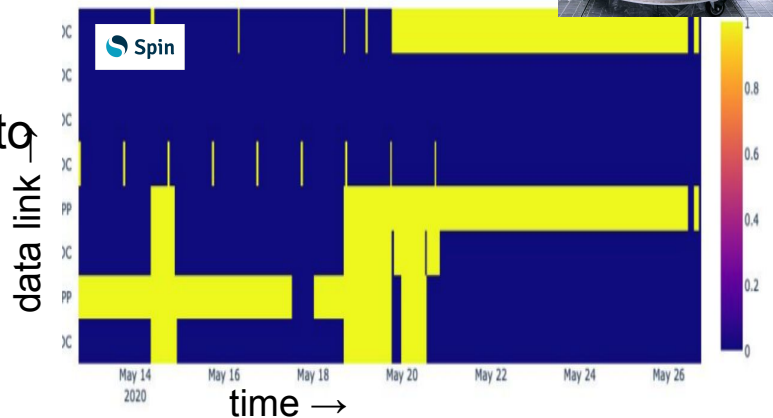
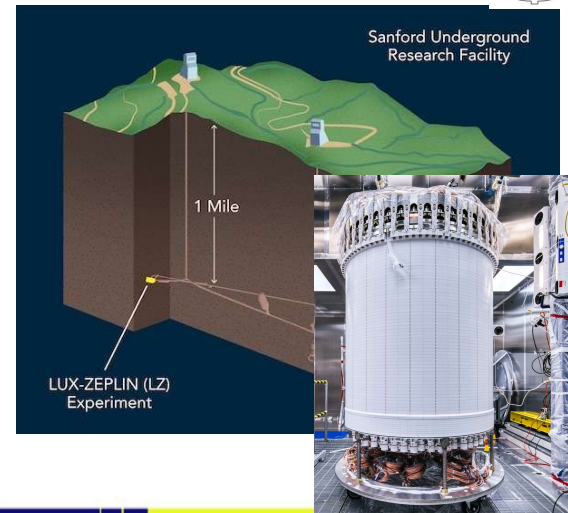
# LZ uses NERSC to watch for dark matter particles



NERSC is also the primary US compute center - used for offline simulation production and analysis, and to monitor data quality 24/7:

1. Bring data to NERSC
2. QA and detector health check
3. Archive data at NERSC
4. Send copy of data to UK data center

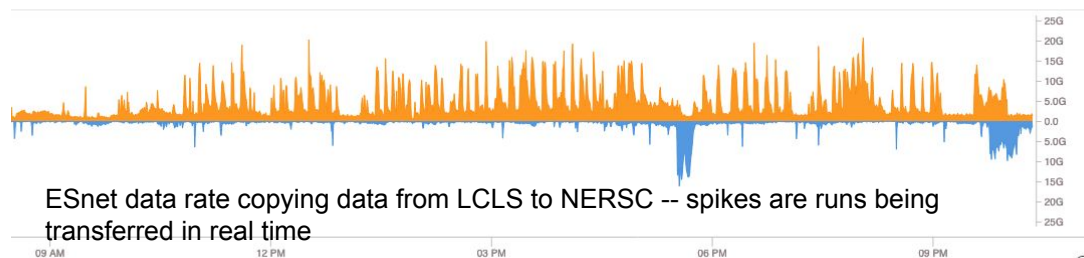
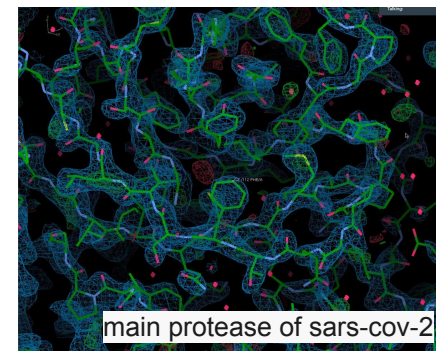
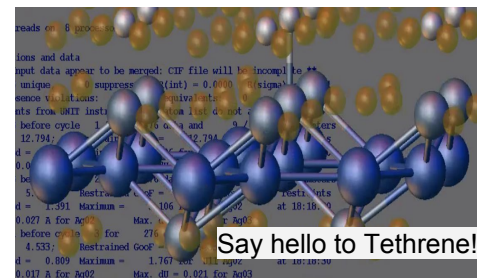
Workflow is operated from NERSC using Spin to coordinate real-time computing to assess data quality and monitor workflow.





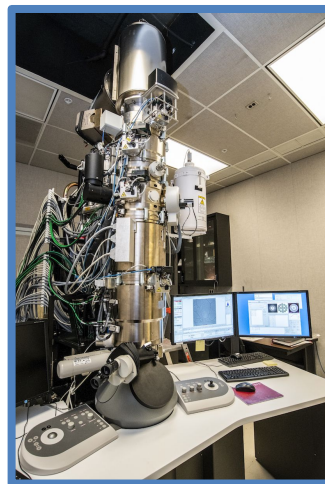
# LCLS is using NERSC for real-time data analysis

- Several experiments at the LCLS (x-ray free electron laser at SLAC) are now using NERSC for real-time data analysis for materials science and Covid-19 research
- Can analyze a 5 minute experiment in ~3 minutes for feedback to beamline staff, transferring 15TB/day to NERSC
  - **Real-time** data analysis using real-time queue and advanced reservations
  - Used services running on **Spin** to orchestrate jobs/parameters/results in real time between several concurrent remote users



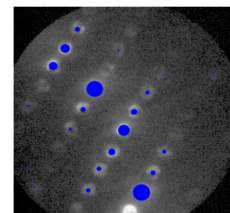
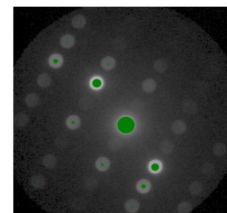
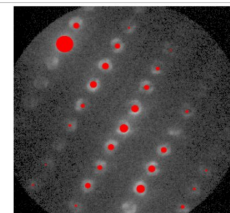
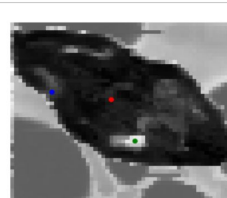
# National Center for Electron Microscopy uses NERSC to process large data sets live during experiments

- NCEM uses Jupyter and Dask for interactive exploration and analysis of EM images
- Dask is a powerful backend to manage remote workers on a cluster via Python notebooks.
- We re-engineered the Dask backend for seamless HPC integration
- Achieved **20-50x speedup** on NCEM Py4DSTEM Notebooks



jupyter  
nbviewer

JUPYTER FAQ



All DPs

```
]: # Get peaks
corrPower = 0.8
sigma = 2
edgeBoundary = 20
maxNumPeaks = 70
minPeakSpacing = 50
minRelativeIntensity = 0.001
verbose = True

peaks = find_Bragg_disks(dc, probe_kernel.data2D,
                        corrPower=corrPower,
                        sigma=sigma,
                        edgeBoundary=edgeBoundary,
                        minRelativeIntensity=minRelativeIntensity,
                        minPeakSpacing=minPeakSpacing,
                        maxNumPeaks=maxNumPeaks,
                        verbose=verbose)
```

	<b>Time-sensitivity</b>	<b>Have Edge Computing?</b>	<b>What do they do when NERSC is unavailable?</b>
<b>ALS</b>	Need NERSC sporadically during a beamline shift	Yes. Small cluster run by IT	Run locally
<b>DESC</b>	Offline data analysis	Yes	Run locally, but no access to collaboration data
<b>DESI</b>	Need NERSC to process nightly telescope data in quasi-realtime	No	Can tolerate some hours delay in analysis Developing alternate sites where workflow can run
<b>JGI</b>	Need NERSC to keep up with sequencers	Building small local capability	Outages result in backlog of data to process - hard to catch up
<b>LCLS</b>	Need NERSC sporadically during experiment shifts	Yes. Local cluster	Run locally with less science insight into the data. Developing alternate sites where workflow can run
<b>LZ</b>	Need NERSC 24/7 during experiment operations	Event detection/signal processing only	No NERSC means cannot run full DAQ Developing alternate sites where workflow can run (v hard)

# Conclusions

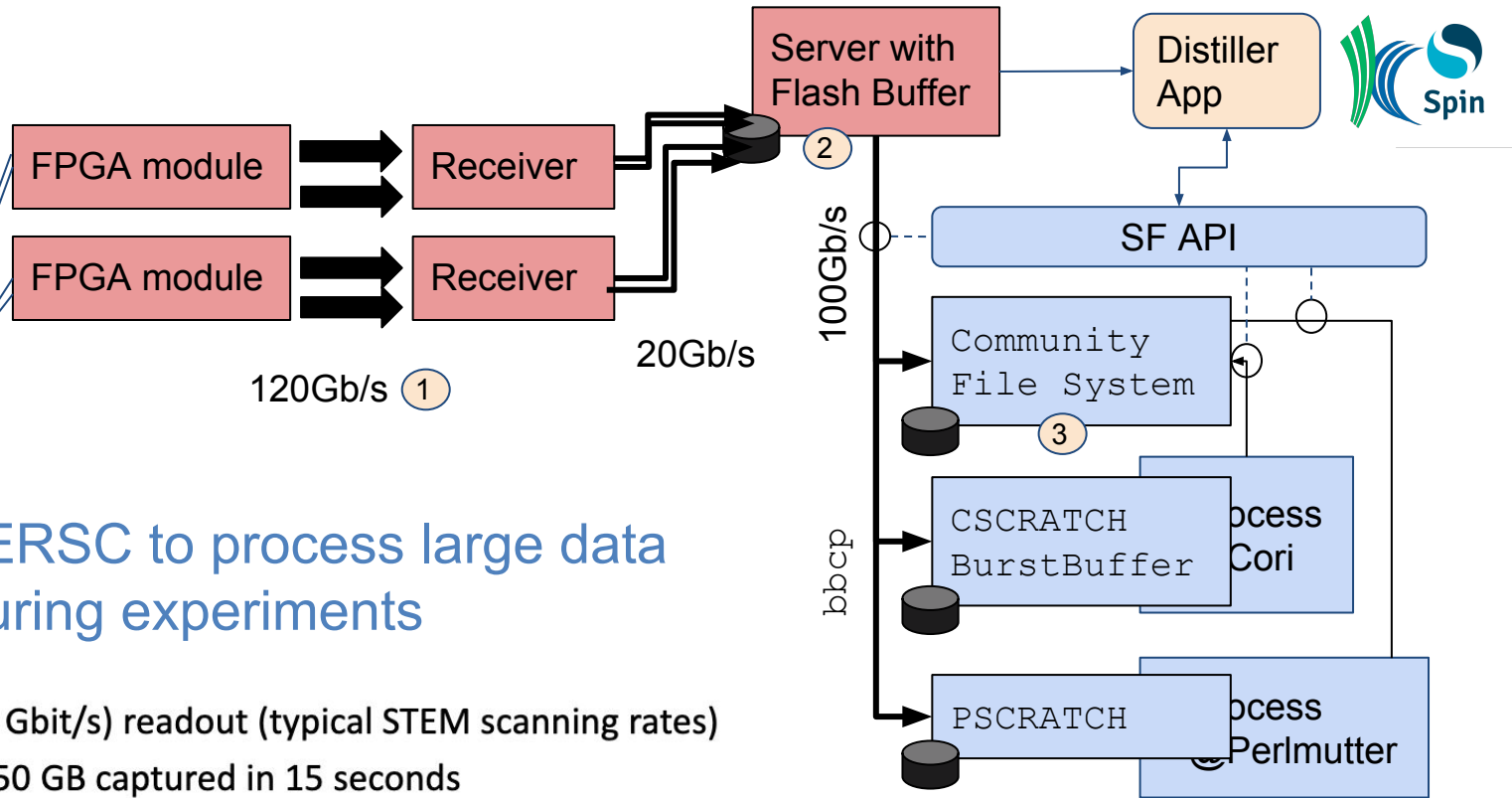
- NERSC has been importing more data than exporting for over a decade, ‘at the edge’
- NERSC *augments* the edge by providing large scale compute - data storage and a suite of tools and services for processing, sharing and collaborating on data
- NERSC recommends experiments have an alternate compute strategy for when NERSC is in maintenance or has system issues – either their own edge computing or an alternate site
- It would be hard to keep Perlmutter busy with purely data analysis from experiment workload - the bandwidth ingest pinch-point is too tight.

Extra





# National Center for Electron Microscopy ...



... uses NERSC to process large data sets live during experiments

- 87,000 Hz (480 Gbit/s) readout (typical STEM scanning rates)
- 1kx1k scan is 650 GB captured in 15 seconds
- Data pipeline: FPGA → RAM → Flash storage → Sparse HDF5

15 sec (1) 140 sec (2) 5 min (3)

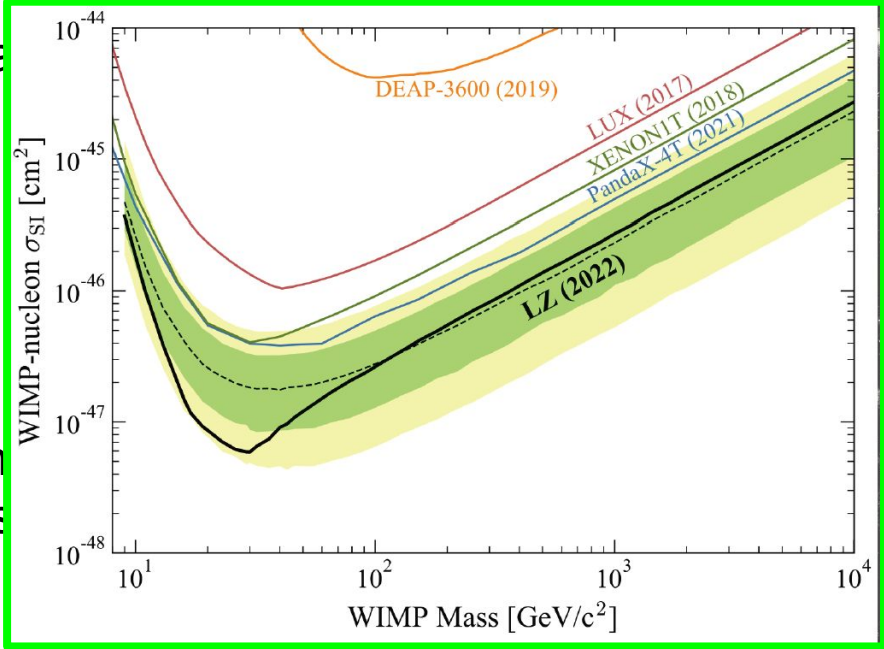


# LZ uses NERSC to watch for dark matter particles

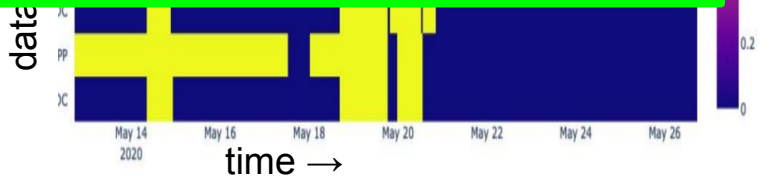


NERSC is the primary US compute center used for offline simulation production and analysis.

First physics results!  
LZ is the most sensitive  
Dark Matter experiment  
currently taking data



quality and monitor workflow.



# The Superfacility Project

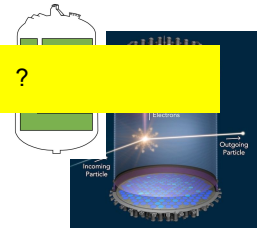
Which of our SF use cases have local edge computing?



Lead by Debbie Bard



?



?

ALS



? They have clusters run by IT, but do some beamlines go right to NERSC for processing?

FIC

Facilities Integrations Collaborations

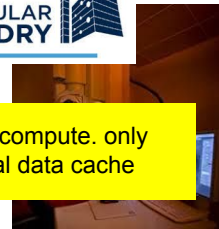
JGI-EMSL Coll

What are JGI's latest capabilities. Did they get IT to run something for them?



?

MOLECULAR FOUNDRY



No compute. only local data cache

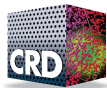
LCLS



Yes



No, which is why they are often grumpy when we go down



AMCR SciData

# The Superfacility Project

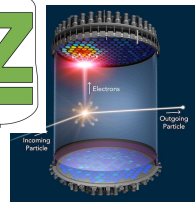
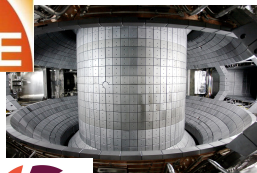
## Goal:

Enable automated pipelines that analyze data from remote facilities at large scale, without routine human intervention, using these capabilities:

- **Real-time** computing support
- Dynamic, high-performance **networking**
- Data management and movement tools, incl. **Globus**
- **API-driven** automation
- HPC-scale notebooks via **Jupyter**
- Authentication using **Federated Identity**
- Container-based edge services supported via **Spin**



Lead by Debbie Bard



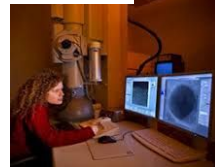
ALS



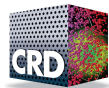
**FICUS**  
Facilities Integrating  
Collaborations for User Science  
JGI-EMSL Collaborative Science Call



**MOLECULAR  
FOUNDRY**



**LCLS**

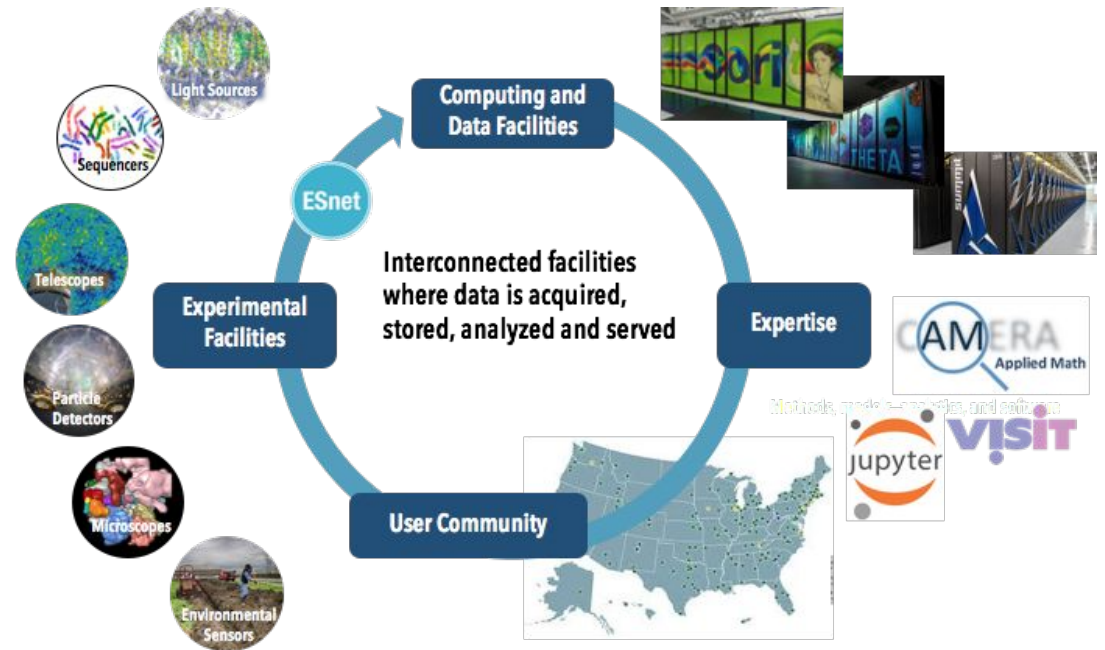


AMCR  
SciData

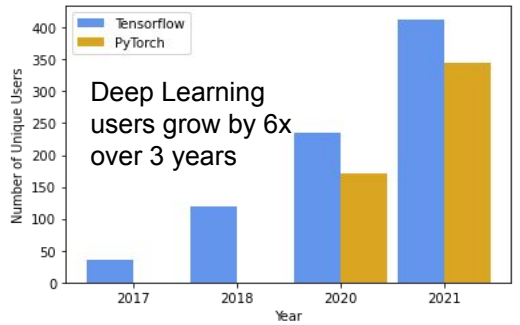
# The Superfacility Model: an ecosystem of connected facilities, software and expertise to enable new modes of discovery

Superfacility@LBNL: NERSC, ESnet, AMCR, & SDD working together to support experimental science

- A model to integrate experimental, computational and networking facilities for reproducible science
- Enabling new discoveries by coupling experimental science with large scale data analysis and simulations

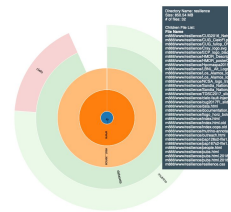


# Capabilities for Experimental Facility Users



Deep Learning users grow by 6x over 3 years

Optimized ML software deployed



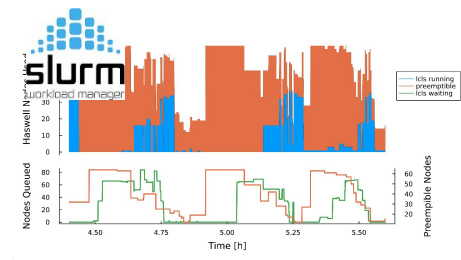
Developed a PI Toolbox for managing data

Choose Your Institution

Recent institutions

- National Energy Research Scientific Computing Center Web Login nersc.gov
- Lawrence Berkeley National Laboratory lbl.gov

Federated Identity



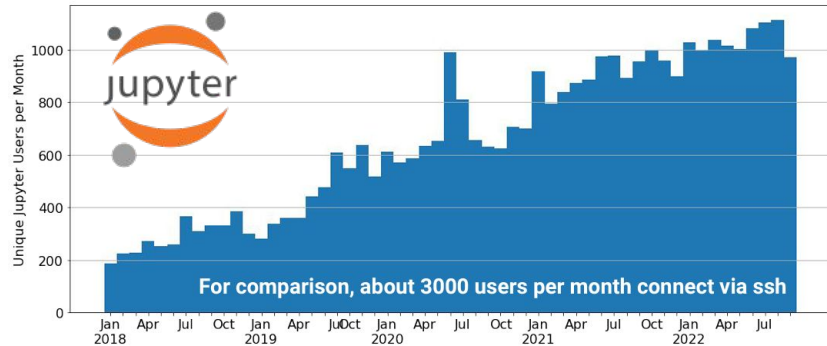
Enabled real-time jobs while maintaining high utilization



Deployed Globus Sharing and End-points for collaborations

**Superfacility API** <sup>1.2</sup>

[ Base URL: /api/v1.2 ]  
/api/v1.2/swagger.json



**Spin**

Deployed an internal cloud for hosting data services and Portals

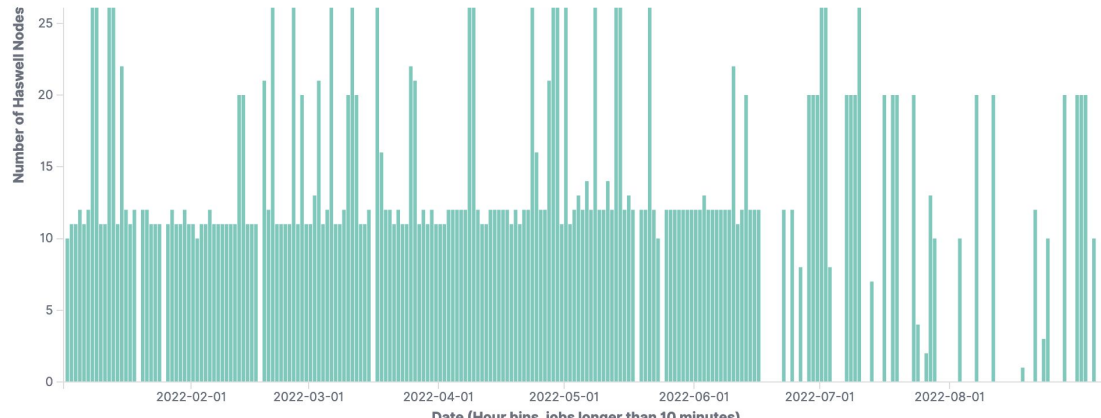
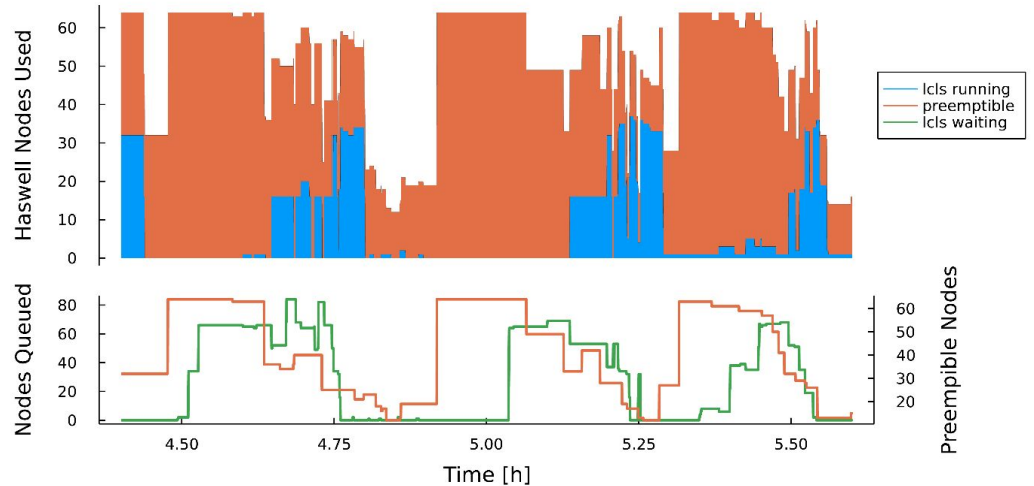


U.S. DEPARTMENT OF **ENERGY**

Office of Science

# NERSC now supports real-time capabilities

- **NERSC has a dedicated pool of real-time nodes for approved projects**
- **NERSC can also support reservations for experiments and enables pre-emptible jobs to keep utilization high**





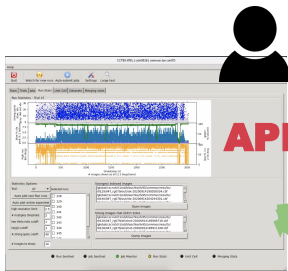
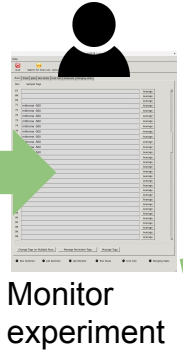
# LCLS is using NERSC for realtime collaborative distributed data analysis

**SLAC** NATIONAL ACCELERATOR LABORATORY



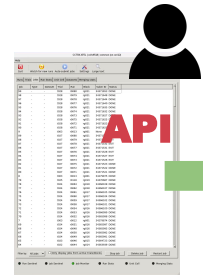
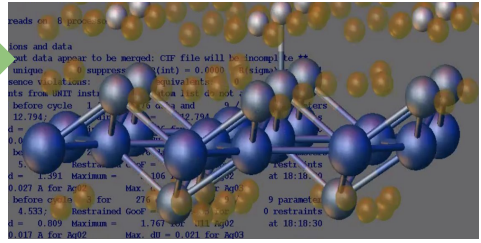
**ESnet**  
ENERGY SCIENCES NETWORK

Incoming data



Monitor analysis

Science!



Submit jobs

cctbx.xfel



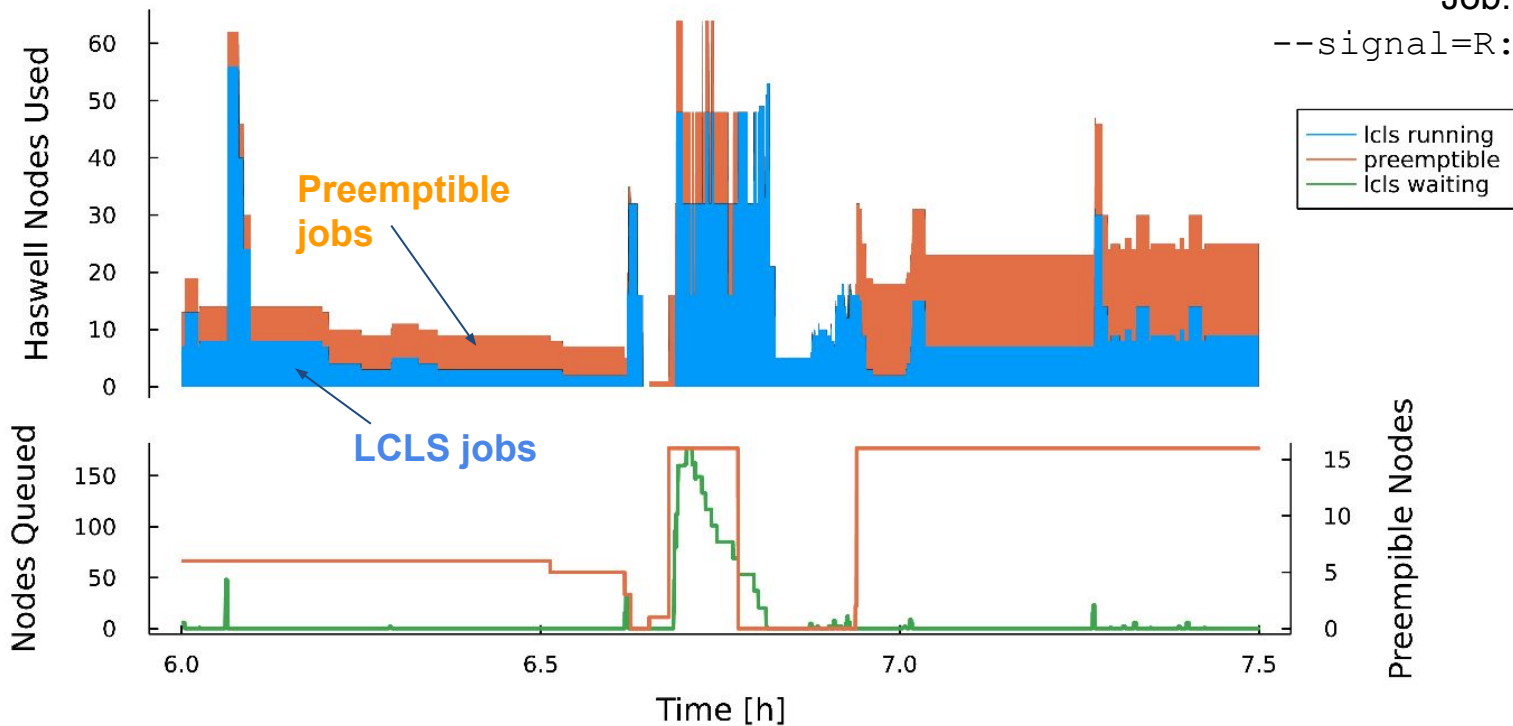
Can analyze a 5 minute experiment in ~3 minutes for feedback to beamline staff, transferring 15TB/day to NERSC

# Avoid Waste with Preemptible jobs that can run in reservations

Reservation: MaxStartDelay=5

Job: #SBATCH

--signal=R:INT@300

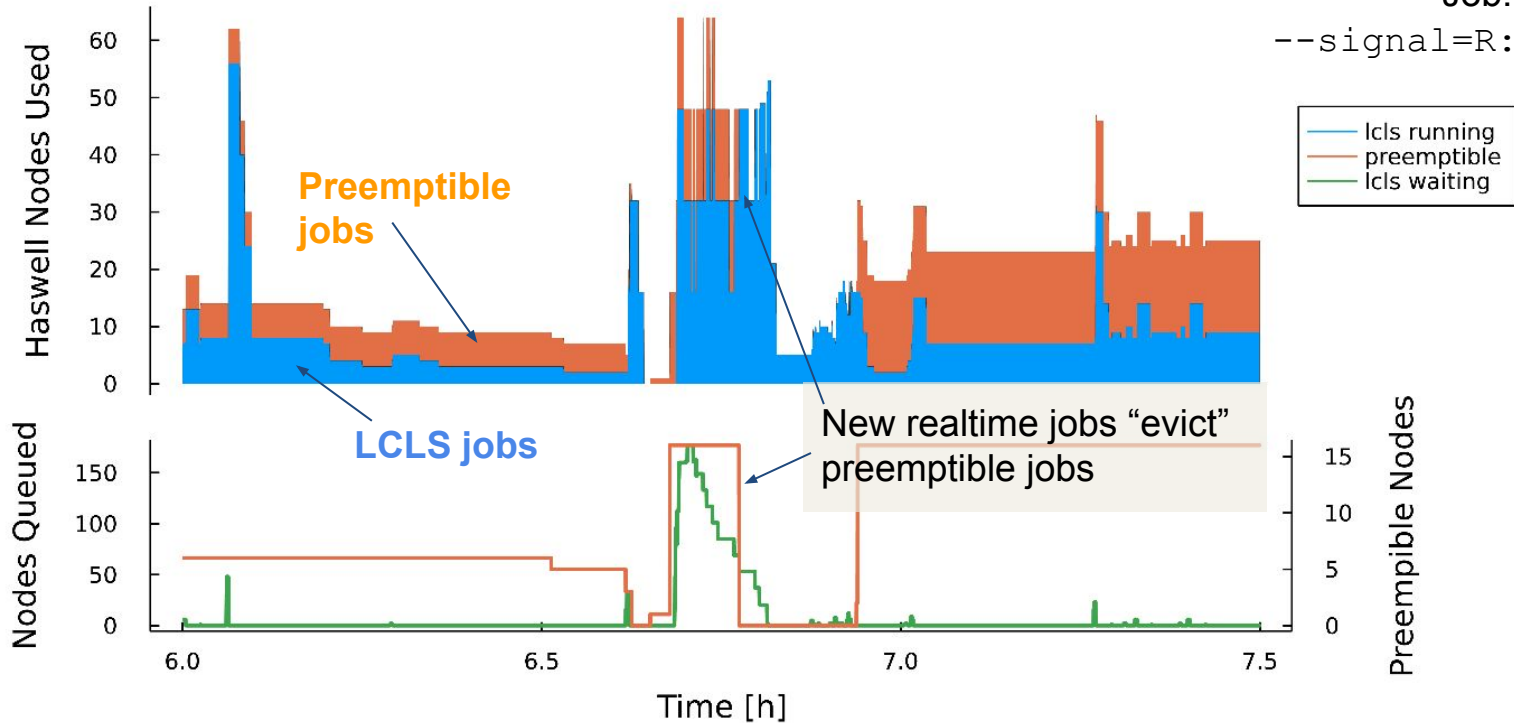


# Avoid Waste with Preemptible Reservations

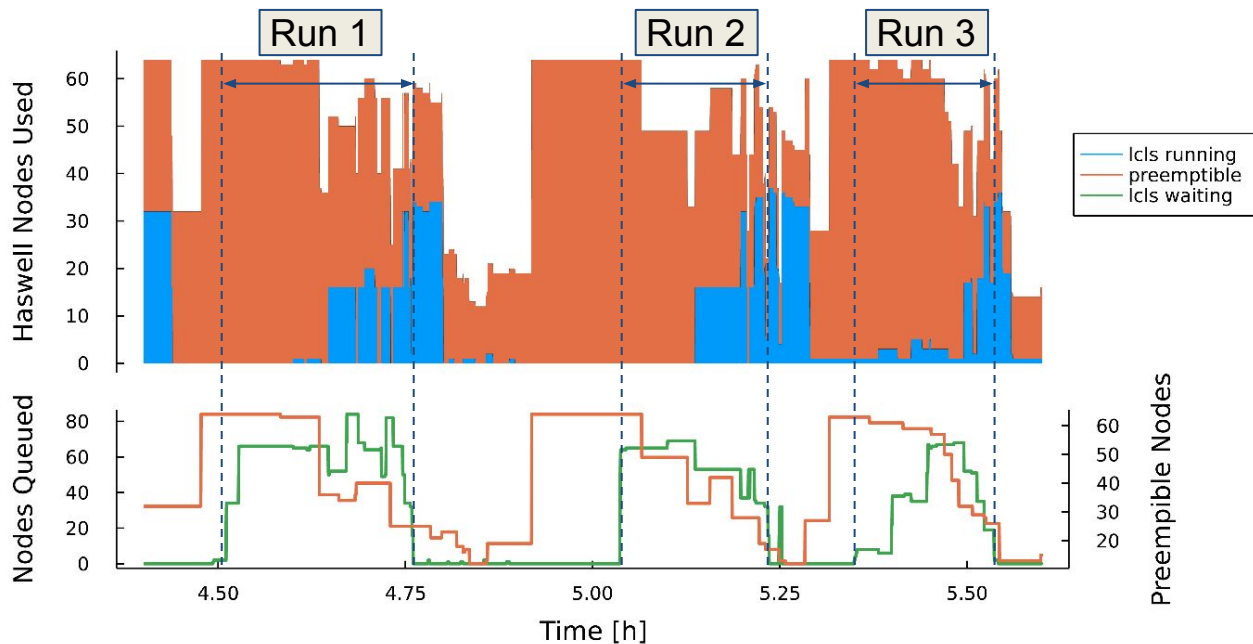
Reservation: MaxStartDelay=5

Job: #SBATCH

--signal=R:INT@300



# Preemption Fills the “Gaps” in Runs



# Spin: Container Services for Science



Many projects need more than HPC.

***Spin is a platform for services.***

Users deploy their **science gateways, workflow managers, databases, and other network services** with Docker containers.

- *Access HPC file systems and networks*
- *Use public or custom software images*
- *Orchestrate complex workflows*
- *Secure, scalable, and managed*



## Some projects using Spin:



Track and compare analyses of nightly sky surveys

science gateway



Classify and store reusable earth sciences data

data repository



Manage production genomic workflows and data at scale

science gateway



Process real-time events for dark matter detection

workflow manager



Explore materials properties or build simulated materials

science gateway



# Jupyter: supercharge interactive supercomputing

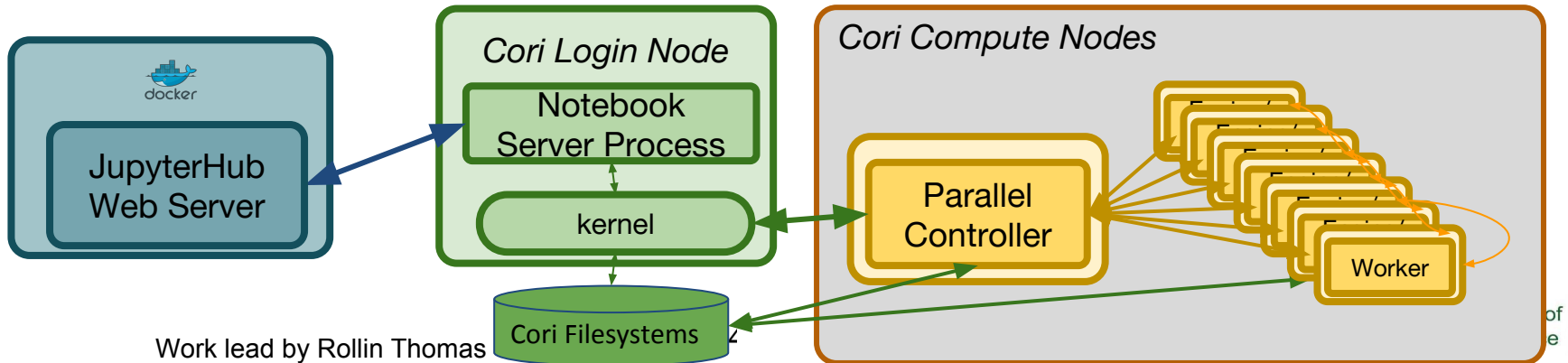
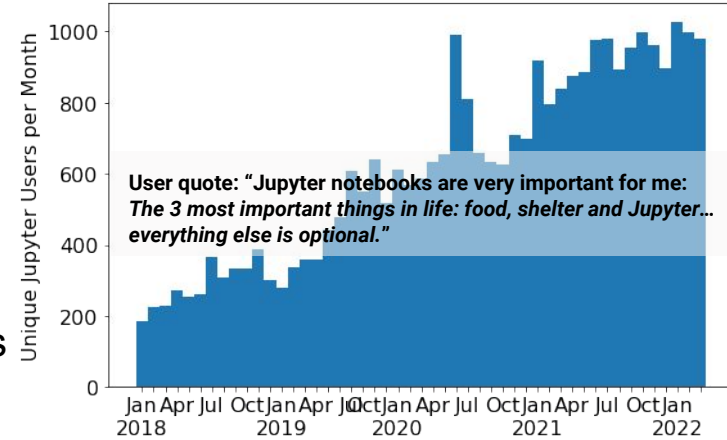


## NERSC leads in HPC-aware Jupyter deployments

- Patterns and frameworks for connecting Jupyter with HPC
- Data analytics/AI platform in an HPC environment
- Interactive visualization and shareable analysis workflows
- Reproducible science through containerization w/SciData Division

## Interactive supercomputing: Jupyter Notebook + HPC Workers

- Launch workers in a short turnaround queue, leveraging our API
- Communicate with distributed analytics clusters (e.g. IPyParallel, Dask)



# Federated Identity (FedID): one identity for many facilities

Users link their home identity to their NERSC account, then use it to log in.




- Simplifies cross-facility workflows
- Users have fewer passwords and login pages
- Home institution manages account lifecycle
- NERSC still manages authorization
- Core technology is established and mature
- *Policy/trust decisions were the bulk of our work*

***Future: DOE DCDE project is building and promoting similar efforts at other sites***

**NERSC** National Energy Research Scientific Computing Center

### Choose Your Institution

Recent Institutions

-  National Energy Research Scientific Computing Center  
nersc.gov
-  Argonne National Laboratory  
anl.gov
-  Oak Ridge National Laboratory  
ornl.gov

+ Add another institution      - Remove an institution

New feature! Users from some national laboratories can now log in to NERSC with credentials from their home institution. [Learn More](#)

# Machine-readable supercomputers: the Superfacility API

**Vision: all NERSC interactions are callable;  
backend tools assist large or complex operations.**

## Endpoints currently deployed:

<code>/meta</code>	information about this Superfacility API installation
<code>/status</code>	NERSC component system health
<code>/account</code>	Get accounting information about the user's projects
<code>/utilities</code>	basic file browsing, upload and download of small files to and from NERSC
<code>/storage</code>	Transfer files between Globus endpoints.
<code>/compute</code>	Run commands and manage batch jobs on NERSC compute
<code>/tasks</code>	Get information about your pending or completed tasks
<code>/reservations</code>	submit and manage future compute reservations

## Superfacility API <sup>1.2</sup>

[ Base URL: /api/v1.2 ]  
/api/v1.2/swagger.json

**SFapi**

API access to NERSC

**meta** information about this Superfacility API installation

GET /meta/changelog

GET /meta/config

**status** NERSC component system health

GET /status

GET /status/notes

GET /status/notes/{name}

GET /status/outages

GET /status/outages/planned

GET /status/outages/planned/{name}

GET /status/outages/{name}

GET /status/{name}

**account** Get accounting information about the user's projects

POST /account/groups

GET /account/groups

36 <https://api.nersc.gov/>

# Bandwidth Pyramid from Perlmutter's point of view



get data from taylor to flip graph so Perlmutter bandwidth is on top

7 Tb/s

This is out 2x 400Gb/s +2 x 100Gb/sec →

Provides connectivity to border plus between systems →

Is this just bandwidth through login infrastructure or does it include FS? →

Is this bi-section bandwidth? →

