HPC and small experiments

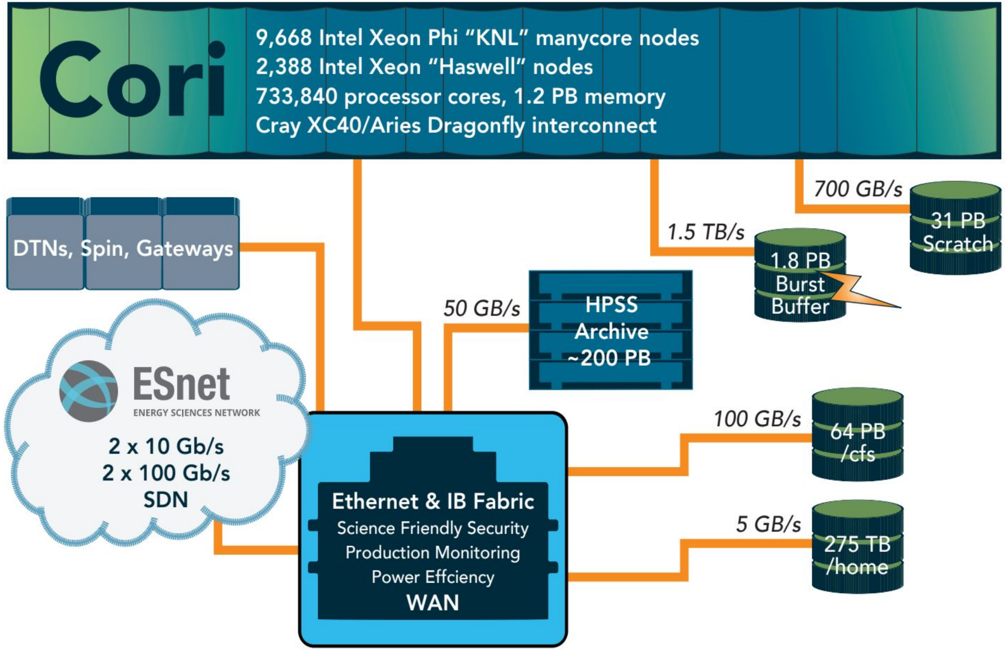
Snowmass Small
Experiments workshop 2021

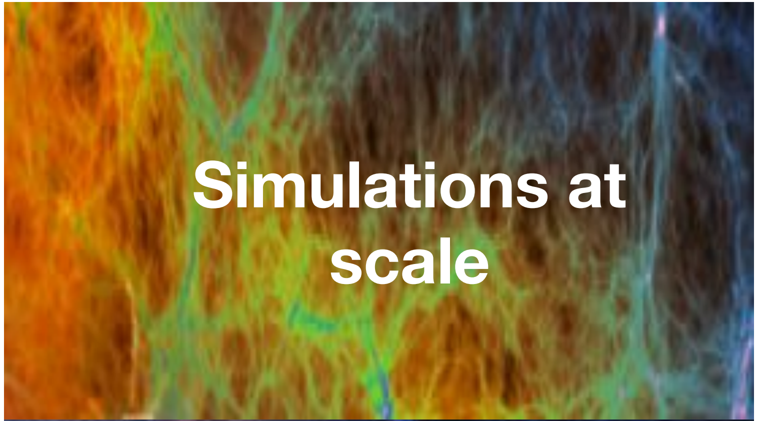
Debbie Bard
Group Lead, Data Science Engagement
NERSC

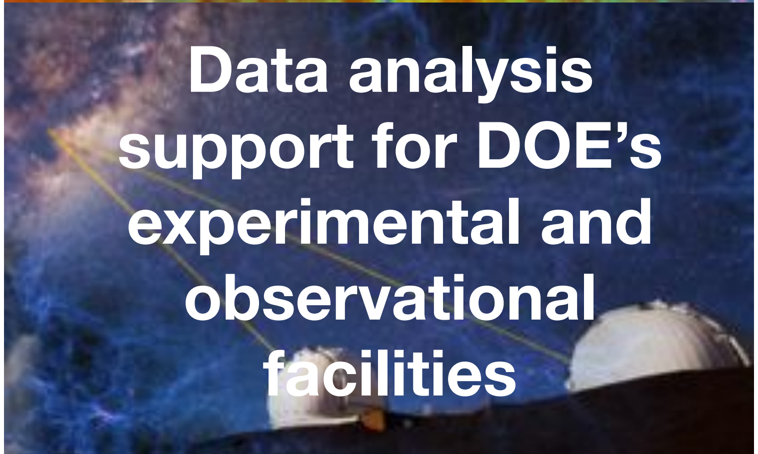
NERSC is the **mission** High Performance Computing and Data facility for the DOE SC

8,000+ Users, 800+ Projects

2000+ NERSC citations per year





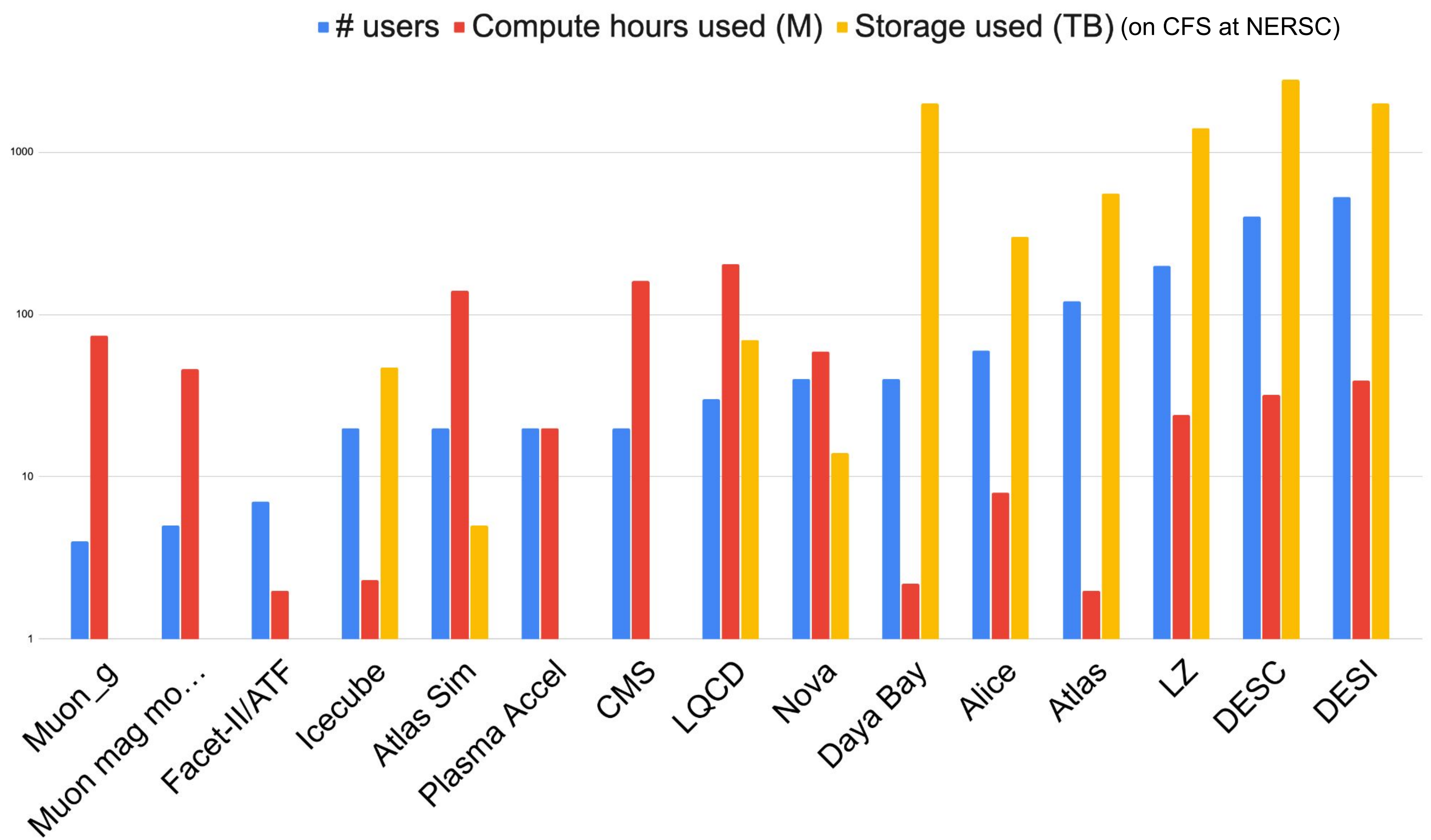


We have a long history of engagements with experimental facilities

Experiments operating now

Future experiments

What is a “small experiment”, from an HPC perspective?



What is a "small experiment", from an HPC perspective?

- From a purely resource-needs perspective, we see very similar stories from most HEP projects (experiment and theory)
  - Similar compute, storage and # users
  - From this perspective, we don't "see" experiment size
- Some of the issues I've heard in this workshop are not unique to "small" experiments, but are exacerbated by the smaller human resources available for computing
  - Getting codes running on unfamiliar architectures
  - System uptime / resiliency
  - Tools that are easy to stand up and maintain

Key challenge from HPC perspective: Porting workflows to new architectures

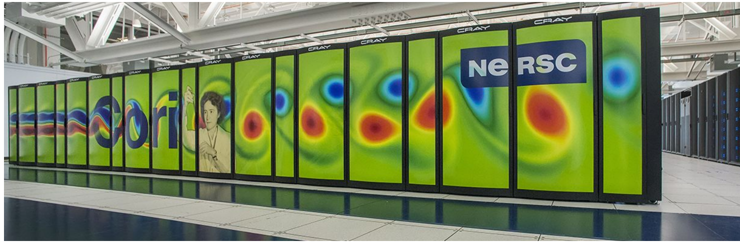
Workflows from experimental facilities include entire ecosystems of hardware, networking, storage, middleware libraries. Are these harder to fit to new architectures, or will they more adaptable?

- **Energy Efficient architectures are here to stay**
  - Edge devices set to provide specialized architectures for specific problems.
- **Storage Technologies are changing**
  - Flattening of storage hierarchy, Lustre on SSDs, object stores
- **Edge services are changing**
  - Increasing use of containers and container orchestration, scalable databases.

These challenges will require a new approach to workflow optimization, not just code optimization.

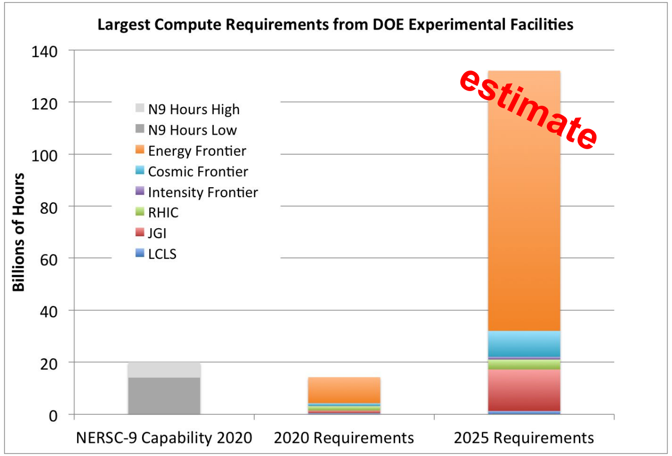
NERSC has a dual mission to advance both science and the state-of-the-art in supercomputing

- **This dual mission is a congressional mandate.**
- We collaborate with computer companies years before a system's delivery to deploy advanced systems with new capabilities at large scale
- We provide a highly customized software and programming environment for science applications
- Our staff provide advanced application and system performance expertise to users



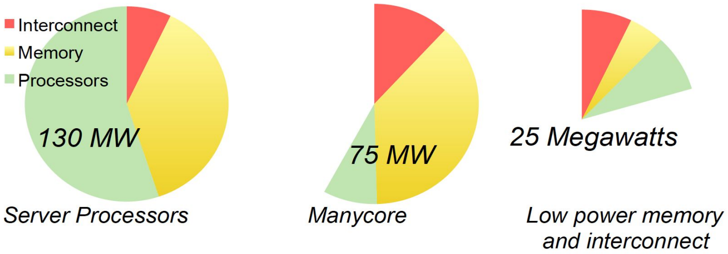
HPC must become more energy efficient

- Scientific compute needs are not getting any smaller
- Pressing question: how can we supply and support these needs?
- **Power/cooling is the main constraint in supercomputer design**
  - Edison: 2.6PFlops, *2.5 MW*, 1 GFlops/Watt
  - Cori: 14 PFlops, *3.6 MW*, 3.5 GFlops/Watt
  - Perlmutter: 70 PFlops, *2.6 MW*, 27.4 GFlops/Watt



TOP500
@top500supercomp

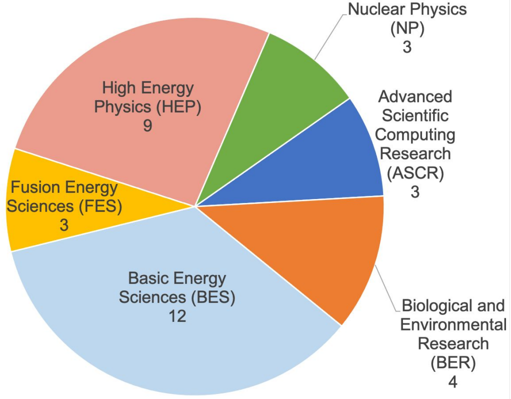
Perlmutter at NERSC/LBNL at No 5 is highest ranked new system It also is #3 on HCG, #4 on HPL-AI and #6 on the Green500!



NERSC

NESAP: NERSC application readiness

- NESAP is NERSC's Application Readiness Program for preparing our workload for new systems.
- Partner with application development teams and vendors to port & optimize key applications of importance to DOE SC.
- Share lessons learned with with NERSC community via documentation and training + Hackathons.
- Resource Available to Teams: NERSC Staff liaisons, performance postdocs, access to vendor application engineers, hackathons, early access to hardware (GPUs on Cori and Perlmutter)
- 3 NESAP programs: **Simulation**, **Data**, **Learning**



Tier 1 NESAP Teams

+29 Tier 2 NESAP teams
58 Total NESAP Teams

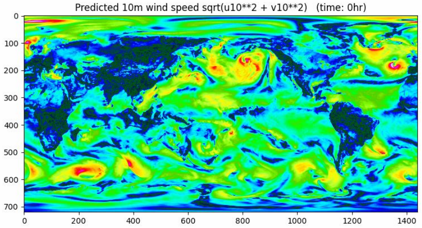
NERSC
BERKELEY LAB Bringing Science Solutions to the World
U.S. DEPARTMENT OF ENERGY Office of Science

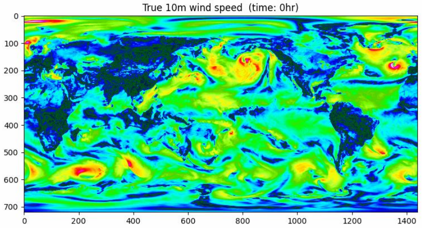
NESAP Successes in Data/Learning

Data-driven Atmospheric modeling

Jaideep Pathak
NESAP Postdoc

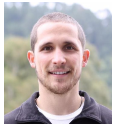
- Prediction of hi…solution atmospheric flow variables
- Using “Fourier Neural Operator” in collaboration with NVIDIA - expected to showcase at GTC
- 2.9x improvement in throughput using Perlmutter A100 compared to Cori V100



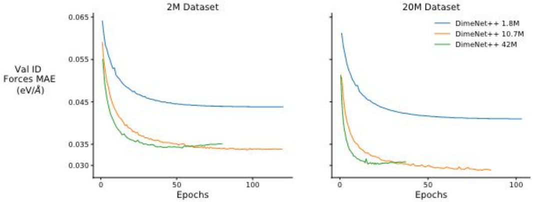


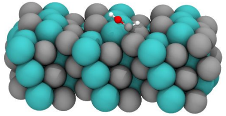
Open Catalyst Project:

Brandon Wood
NESAP Postdoc



- GraphNNs to accelerate catalyst discovery for energy storage and climate change mitigation
- [NeurIPS 2021 Competition]
- Collaboration with CMU + Facebook
- PM allows bigger models - 1000s GPUs

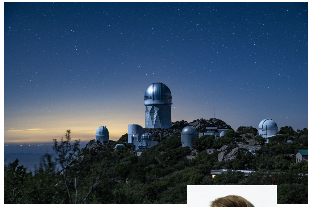


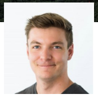


[https://opencatalystproject.org/]

DESI

- DESI Spectral Extraction is an image processing code implemented in Python.
- **2.5x** improvement in per-node throughput using Perlmutter A100 compared to Cori V100 GPU (x25 compared to Edison).

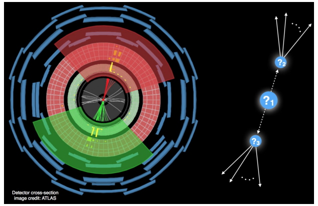




Daniel Margala
(NESAP Postdoc)

Anomaly detection, unfolding and fast simulation in particle physics

- Deep learning techniques now used for [published searches for fundamental particles at the LHC], [production fast-simulation] and “[unfolding]”
- Expanding to more complex models/approaches and higher-fidelity generative networks
- Collaboration with LBL Physics Division (Ben Nachman and others)



Vinicius Mikuni
NESAP Postdoc

Key challenge from HPC perspective: policy optimization

The biggest challenges to success in this area are not necessarily technical.

Simple technical needs are often the most difficult to implement due to policy/security considerations, e.g. federated ID.

- Supporting experiment facility workflows will mean re-thinking the way we assess success
  - Is system utilization the right metric for self-evaluation?
- How can we support workflow resiliency for experiment facilities?
- How can we support real-time workloads while maintaining quality of service for other NERSC users?
- How can we support new categories of users?

These challenges will require new policies for HPC centers

Resiliency

- Several experiments use NERSC for their everyday data processing
  - They have resiliency plans in the event of an outage, but if we can keep a minimal set of services running during planned maintenances we can help them stay productive
  - Data access a priority
- NERSC is addressing workflow resiliency in several ways:
  - Through careful prioritisation and investment, we are able to keep most infrastructure up during outages using generators (power work, scheduled maintenances…)
  - Helping science teams develop more resilient workflows
  - Exploring how to make infrastructure more portable across computing sites through
    - ALCC project (NERSC/ALCF/OLCF)
    - LBNL LDRD (NERSC/Lab cluster)
    - Nascent SC/ASCR effort to develop a blueprint for integrated research infrastructure.

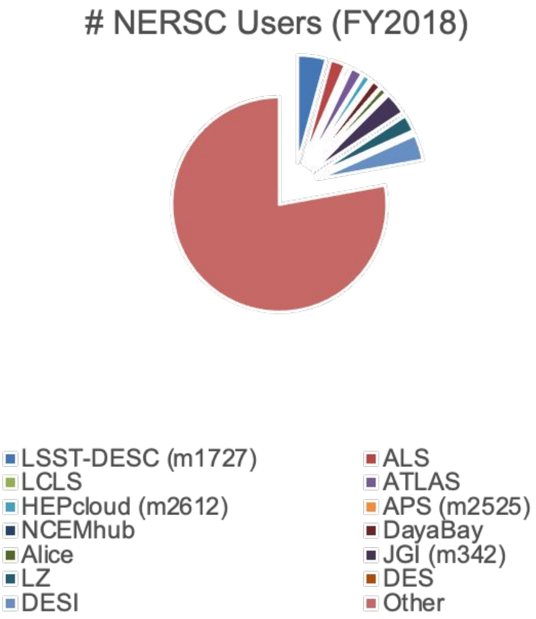
We are now able to keep most of our infrastructure up during power work or routine maintenances

	Network	Spin	Login node	DTN	/global/common (software install filesystem & home dirs)	CFS	HPSS	Compute nodes	Jupyter (needs Spin + NGF, login node)
ALS	✓✓✓	✓✓✓	✓✓			✓✓✓	✓✓	✓	✓
DESC	✓✓✓		✓✓	✓✓✓	✓✓✓	✓✓✓			✓
DESI	✓✓✓	✓✓✓	✓	✓✓✓	✓✓✓	✓✓✓	✓✓	✓	✓
LZ	✓✓✓	✓✓✓	✓	✓✓✓		✓✓✓	✓✓	✓	
JGI	✓✓✓	✓✓✓ (+VMs)		✓✓✓	✓✓✓ (+DNA, seqFS)	✓✓	✓✓	✓✓	

✓✓✓: minimum to keep user productive and basic data access for users
✓✓: Important functionality, but easier to work around
✓: Desirable, gives quasi-full functionality

Key challenge from HPC perspective: scaling user support

- The proportion of NERSC users from experiment facilities has increased over the past 5 years
  - More projects, and more users per project
  - We have >8000 users today. What do we do when we hit 20,000 users? 50,000?
- Staffing not growing proportionally
- How to scale support for EOS facility users?
- Who do the experiments turn to when they run into a problem?

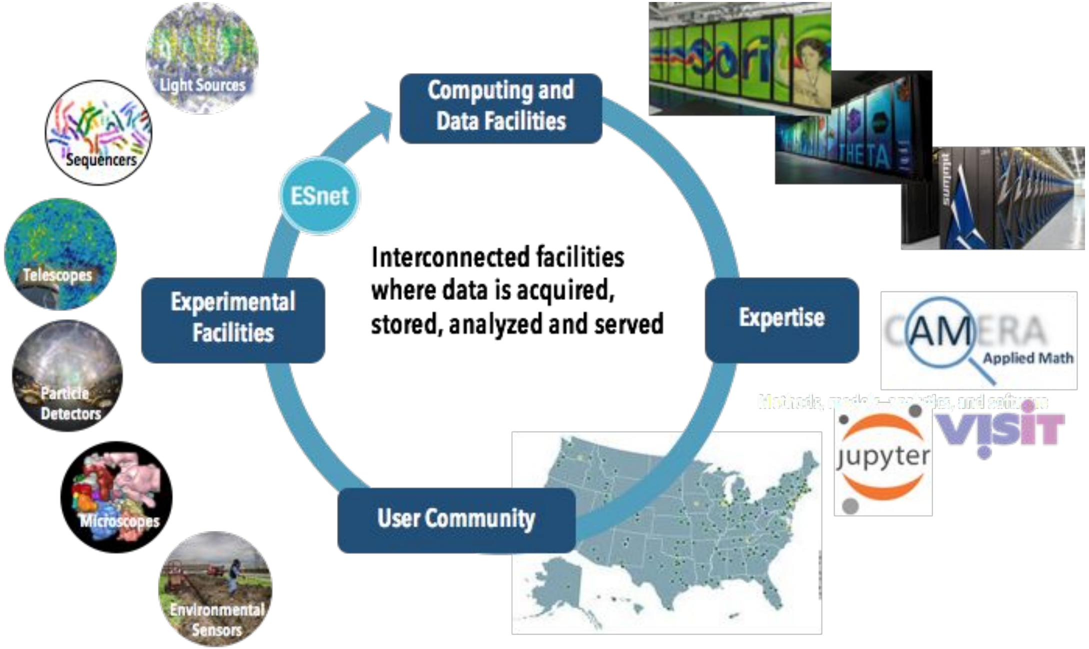


These challenges will require a new approach to user support

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

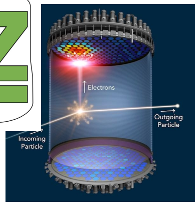
Superfacility@ LBNL: NERSC, ESnet and CRD working together to support experimental science

- 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



NERSC ESnet ENERGY SCIENCES NETWORK CRD COMPUTATIONAL RESEARCH DIVISION BERKELEY LAB Bringing Science Solutions to the World U.S. DEPARTMENT OF ENERGY Office of Science

The LBNL Superfacility ‘project’ coordinates work in CS to support the Superfacility Model



Project Goal:

By the end of CY 2021, 3 (or more) of our 7 science application engagements will demonstrate 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**

ALS

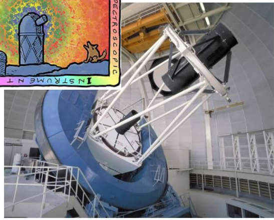
FICUS
Facilities Integrating Collaborations for User Science
JGI-EMSL Collaborative Science Call

LSST DESC
Dark Energy Science Collaboration

MOLECULAR FOUNDRY

LCLS

DARK ENERGY SPECTROSCOPIC INSTRUMENT



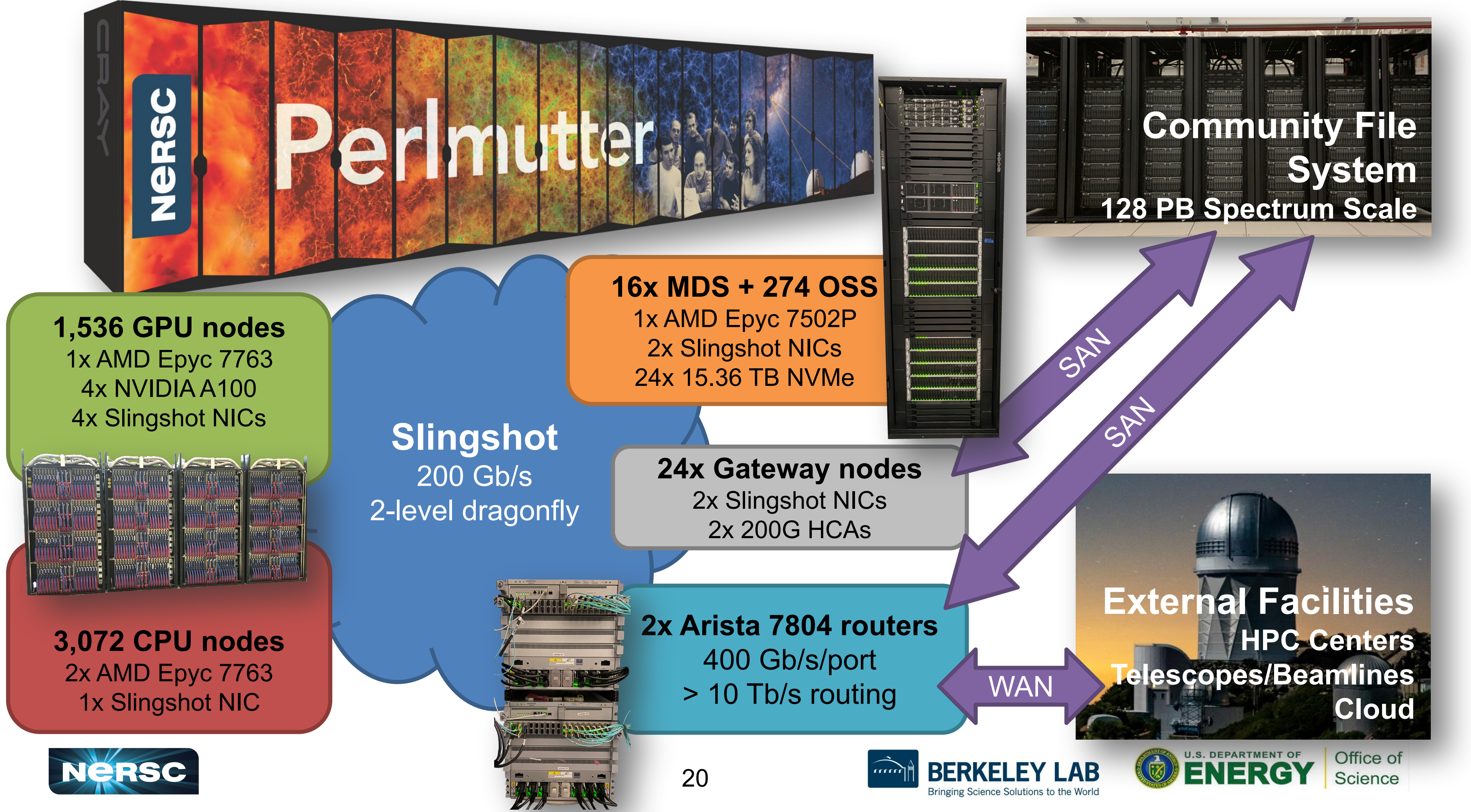
The principles behind our project approach: **integrated**, **scalable**, **sustainable**

- Leverage and **integrate** work being done across many independent teams at LBNL
- Take requirements from multiple user teams
  - No one-off solutions!
  - Scale support to full NERSC user base
  - Iterate with deeply engaged science teams to get the design right
    - detailed supervised surveys, beta testers…
- Use existing, open source, industry standard tools wherever possible
  - Don't want to waste staff time re-inventing the wheel
  - Need to support this workload long-term - cannot rely on custom code only one person understands.

HPC and ~~sm~~all experiments

- From an HPC center perspective, the size of the experiment doesn't really matter. We see similar resource needs from all HEP experiments (and indeed simulation projects).
- Scaling is an issue on the HPC facility side as well
  - New architectures - NESAP
  - More resiliency - policy and investment
  - Scaling - multi-use tools and user support
- Snowmass is a great opportunity to advocate for increased investment and attention
  - Basically in all the areas Stephen said :)

Thanks!



NERSC
BERKELEY LAB
Bringing Science Solutions to the World
U.S. DEPARTMENT OF ENERGY
Office of Science

NERSC has a large and diverse workload

Snapshot of live computing last week:

- 2,062 jobs running
- 10,874 jobs queued
- 95% utilization
- Mixture of simulation and data analysis

Center Overview

Cori Cray XC40
Peak TFlops/s: 30,316
Jobs running: 2,062
Jobs queued: 10,874
Nodes in use: 11,522 (95%)

Storage Usage
Community File System: 35 PB
Cori Scratch: 20 PB
Archive: 214 PB

Current Node Usage Breakdown

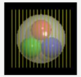
Nodes in Use by Office:

- Basic Energy Sciences: 2,805
- High Energy Physics: 2,770
- Biological & Environmental Research: 854
- Fusion Energy Sciences: 577
- Nuclear Physics: 272
- ALCC: 232
- Other Science Categories: 59

Nodes in Use by Science Category:

- Cosmology/Astrophysics: 1,612
- Physics: 1,230
- Chemistry: 871
- Fusion/Plasma: 514
- Material/Condensed: 502
- Electronic Structure: 487
- Other Science Categories: 2,353

Top Jobs

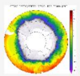


Quantum Chromodynamics with four flavors of dynamical quarks
Office: High Energy Physics
Investigator: Doug Toussaint
Science Area: Physics
Nodes: 1,024 (Cori)
Core Hours Used: 35,070.5

quantum mechanical design of sustainable aliphatic polymers with high glass-transition-temperatures
Office: Basic Energy Sciences
Investigator: Chengyan Zhan
Science Area: Quantum
Nodes: 256 (Cori)
Core Hours Used: 2,368.8

Detector Simulation of the ATLAS Detector on NERSC HPCs
Office: High Energy Physics
Investigator: Paolo Calafiura
Science Area: Physics
Nodes: 200 (Cori)
Core Hours Used: 4,369.7

a new approach to the interacting phonon problem
Office: Basic Energy Sciences
Investigator: Chris Marianetti
Science Area: Structural
Nodes: 192 (Cori)
Core Hours Used: 153,006



Ocean and Sea Ice Interactions around Greenland and the West Antarctic Peninsula in Forced Fine- Resolution Global Simulations
Office: Biological & Environmental Research
Investigator: Julie L. Mcclean
Science Area: Oceanic
Nodes: 176 (Cori)
Core Hours Used: 1,534.7

Requirements reviews and users from experimental facilities describe numerous pain points

- **Workflows** require manual intervention and custom implementations
- Difficult to surge experimental pipelines at HPC facility in '**real-time**'
- I/O performance, storage space and access methods for **large datasets** remain a challenge
- Searching, publishing and sharing **data** are difficult
- **Analysis codes** need to be adapted to advanced architectures
- Lack of **scalable analytics software**

Technical

- **Resilience strategy** needed for fast-turnaround analysis
  - including: coordinating maintenances, fault tolerant pipelines, rolling upgrades, alternative compute facilities...
- No **federated identity** between experimental facilities and NERSC
- Not all scientists want command-line access.

Policy

What features do these experiments need from HPC facilities?

Large volumes of data storage

Scalable IO libraries

More compute hours

Scalable analytics codes

Deadline computing

Resilient workflows to run across multiple computing resources

Interactive access and Jupyter

Edge services for databases, web services, workflow managers...

IO patterns with small/random reads/writes

Sharing data via web portals:

1. **real-time feedback of analysis results**
2. **access to archived data**

Dedicated resources for pipeline/workflow management

High data transfer rates into, out of and within the supercomputer

Co-scheduling compute with experiments

Automate everything!

Spin: Container Services for Science

Spin

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*

docker RANCHER kubernetes

Some projects using Spin:

Project	Description	Service type
Dark Energy Spectroscopic Instrument	Track and compare analyses of nightly sky surveys	science gateway
ESS-DIVE	Classify and store reusable earth sciences data	data repository
JGI Joint Genome Institute, Department of Energy	Manage production genomic workflows and data at scale	science gateway
LZ	Process real-time events for dark matter detection	workflow manager
The Materials Project, materialsproject.org	Explore materials properties or build simulated materials	science gateway

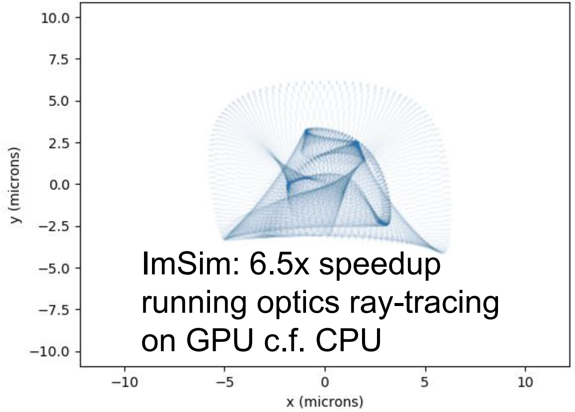
DESC is using NERSC to serve data to the public

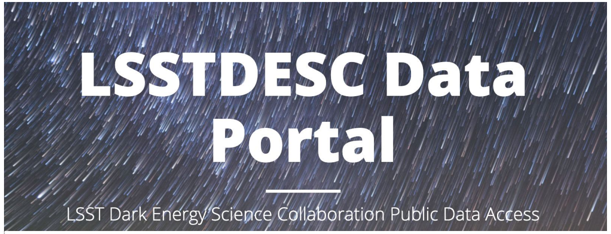
VERA C. RUBIN OBSERVATORY

LSST DESC Dark Energy Science Collaboration

LSST Legacy Survey of Space and Time

- LSST Dark Energy Science Collaboration is using NERSC to produce simulated data
  - Used Spin for associated databases and workflow control for simulation production
  - Analysing it in the collaboration via specialised Jupyter notebooks, scaled out using dask
- Modern Research Data Portal (MRDP) was set up to share data to wider community
  - Based on science DMZ model - data transfer request is outsourced to DTN via Globus. Can handle 3TB/hr
  - easy to stand up using template from Globus
- Managing data for ~500 NERSC users with one of largest CFS allocations
  - Using PI and data dashboards, globus collab tools





https://lsstdesc-portal.nersc.gov/

Machine-readable supercomputers: the Superfacility API

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

Endpoints currently deployed:

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

https://api.nersc.gov/

Devs expect modern modes of access: REST API endpoints, JSON payloads, web auth tokens

- Less user/staff DIY: simpler, standardized tooling (Python, etc)
  - Stable refactor target for established projects
  - Easier on-ramp for new projects
- Fit (not fight) standard software design patterns
  - Shared libraries and API calls
  - Authentication and security models built on OAuth2 Standard and JSON Web Tokens (JWTs)

using the API from a Jupyter notebook to check Cori status

Before we start any computing, let's check whether Cori is up.

```
[3]: health_cori = api("health/resource_statuses/cori", data={"notes":"false", "outages":"true"}, as_form=True)[0]
     print("Cori is %s" % health_cori['status'])
```

```
Cori is active
```

We can also take a look into the future to better plan our work around planned outages.

```
[4]: planned_outages = [o for o in health_cori['outages'] if o['status'].lower()=='planned']
     print(planned_outages)  #make this nicer
```

[{'startdate': '2020-05-20T05:00:00', 'enddate': '2020-05-20T19:00:00', 'description': 'Scheduled Maintenance', 'notes': 'ExVivo and CGPU resources will be unavailab
le during this maintenance.', 'status': 'Planned', 'swo': 'true', 'identifier': 'QXg7SbWP3KAeG0mwkyQS', 'updatedate': None}]

jupyter SFapi

https://docs-dev.nersc.gov/sfapi/

Opening up NERSC to API calls took careful consideration

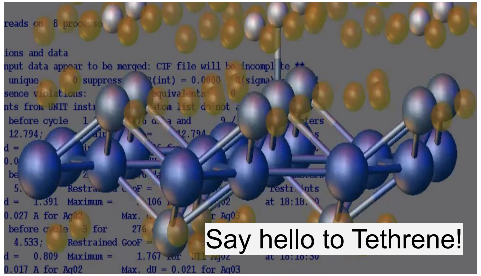
- Conducted UX review
  - An analysis from the user point of view → made changes for functionality and ease of use
- Took inspiration from existing APIs
  - Restructured endpoints to align with scheme from CSCS's FirecREST APi
- Conducted full security review
  - Included both API architecture and new OpenID-based authentication
  - Authentication model requires strict credential lifetimes - need to enforce MFA

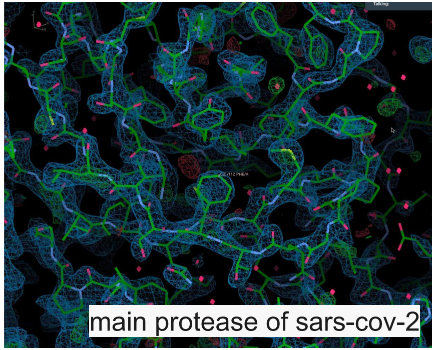
Read-only endpoints:
6 month token lifetime, limited to registered IP addresses

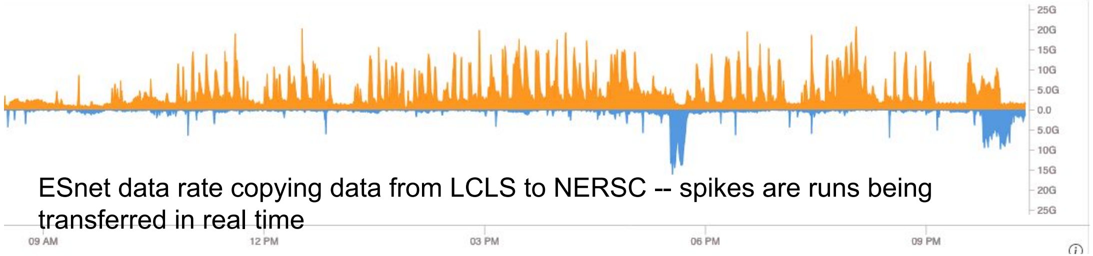
R/W/X endpoints available by request only:
30 day token lifetime, limited to registered IP addresses
Request granted based on specific, defined criteria: Need well defined and credible security and use plan

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 orchestrated jobs/parameters/results in real time between several concurrent remote users


Say hello to Tethrene!

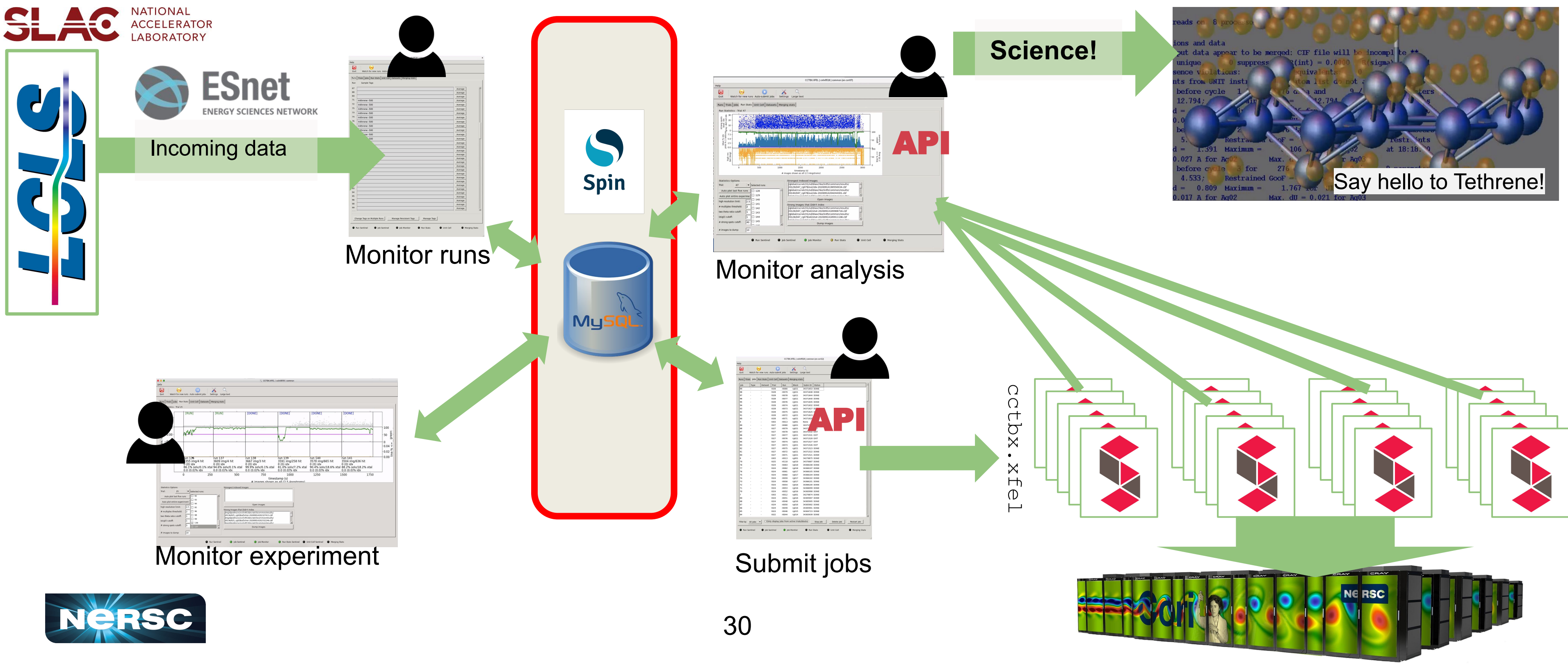

main protease of sars-cov-2



ESnet data rate copying data from LCLS to NERSC -- spikes are runs being transferred in real time

NERSC ESnet ENERGY SCIENCES NETWORK CRD COMPUTATIONAL RESEARCH DIVISION

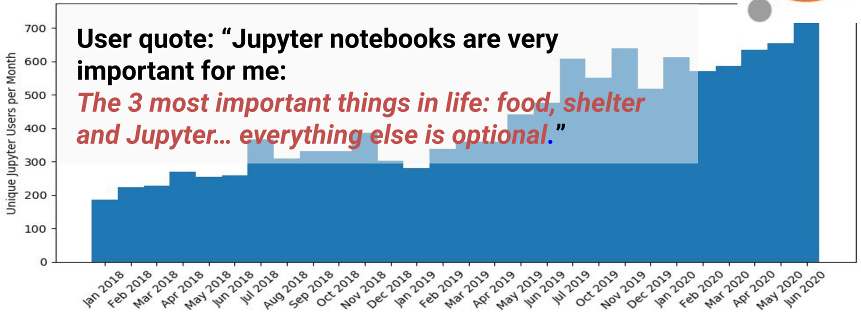
LCLS is using NERSC for collaborative distributed Data Analysis with Spin and the SFapi



Jupyter: supercharge interactive supercomputing

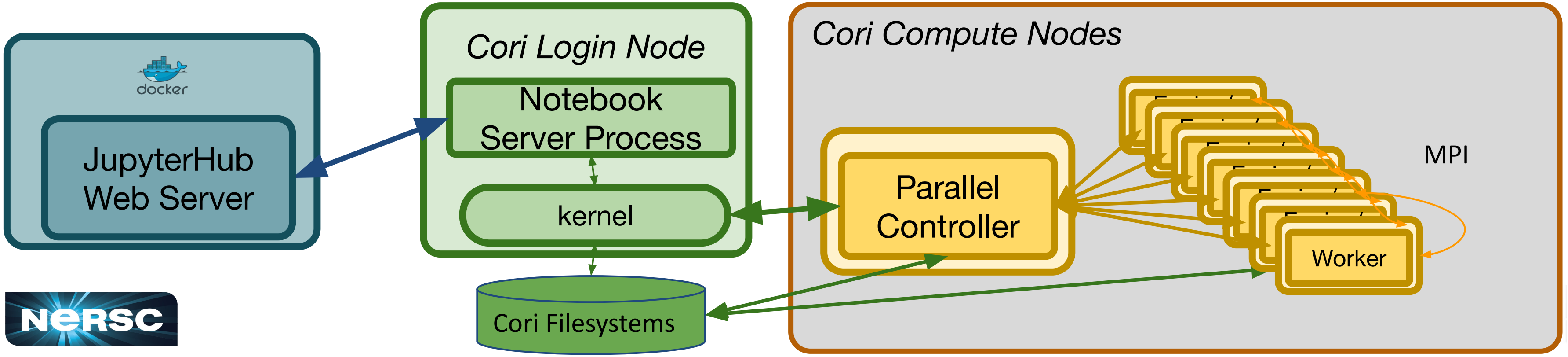
We have deployed an HPC-aware Jupyter service:

- Patterns and frameworks for connecting Jupyter with HPC
- Data Management tools in an HPC environment
- Interactive Visualization
- Reproducible Science through Containerization

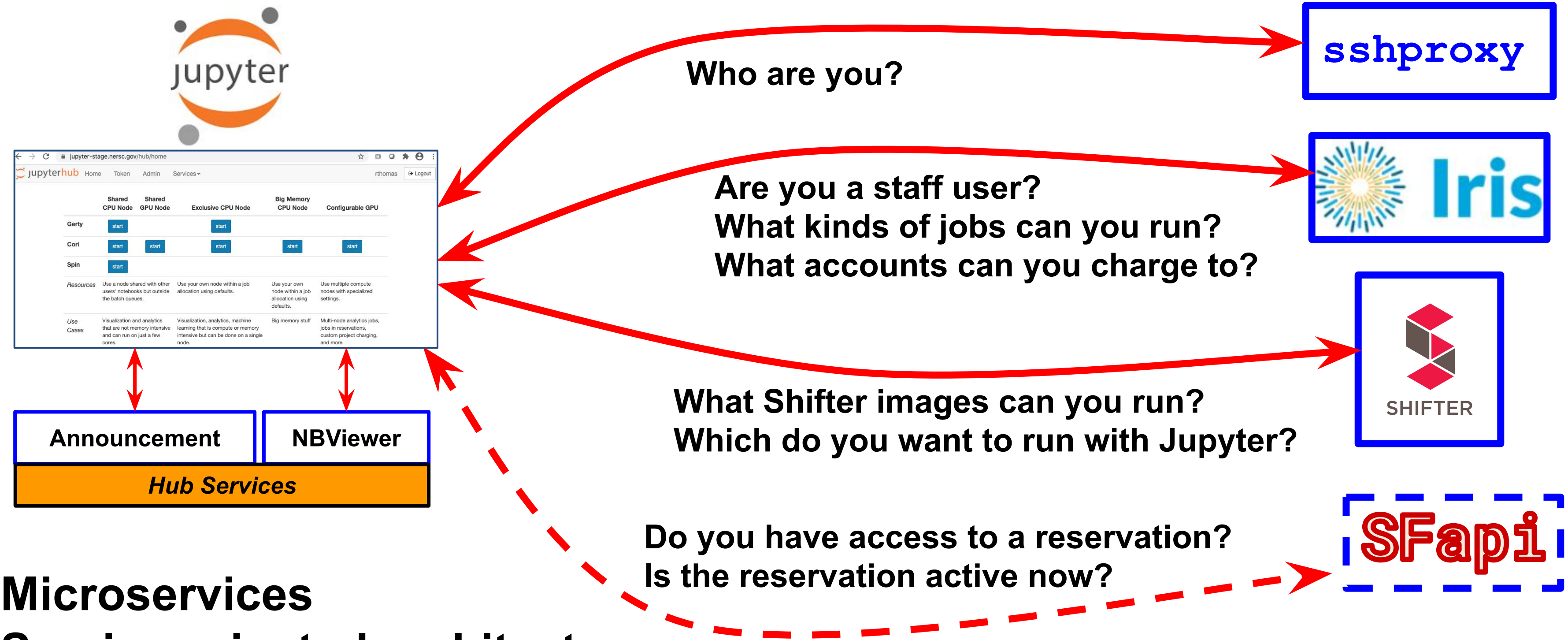


Interactive supercomputing: Jupyter Notebook + HPC Workers

- Launch workers in a short turnaround queue
- Pull results from running HPC Jobs in realtime



Our Hub Leverages NERSC Service APIs

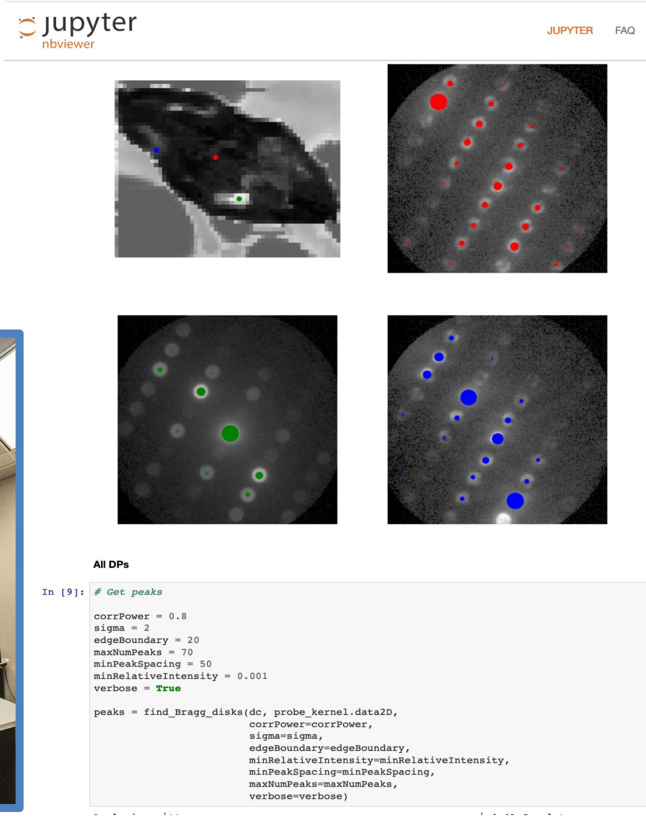


Microservices

Service-oriented architecture

NCEM is using 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
  - Dask integration with Jupyter is not ideal for MPI -based HPC environments , eg no Support for multiple kernels



- National Center for Electron Microscopy: Serial processing of 4D image arrays in numpy - Parallelize it!
- Achieved **20-50x speedup** on NCEM Py4DSTEM Notebooks