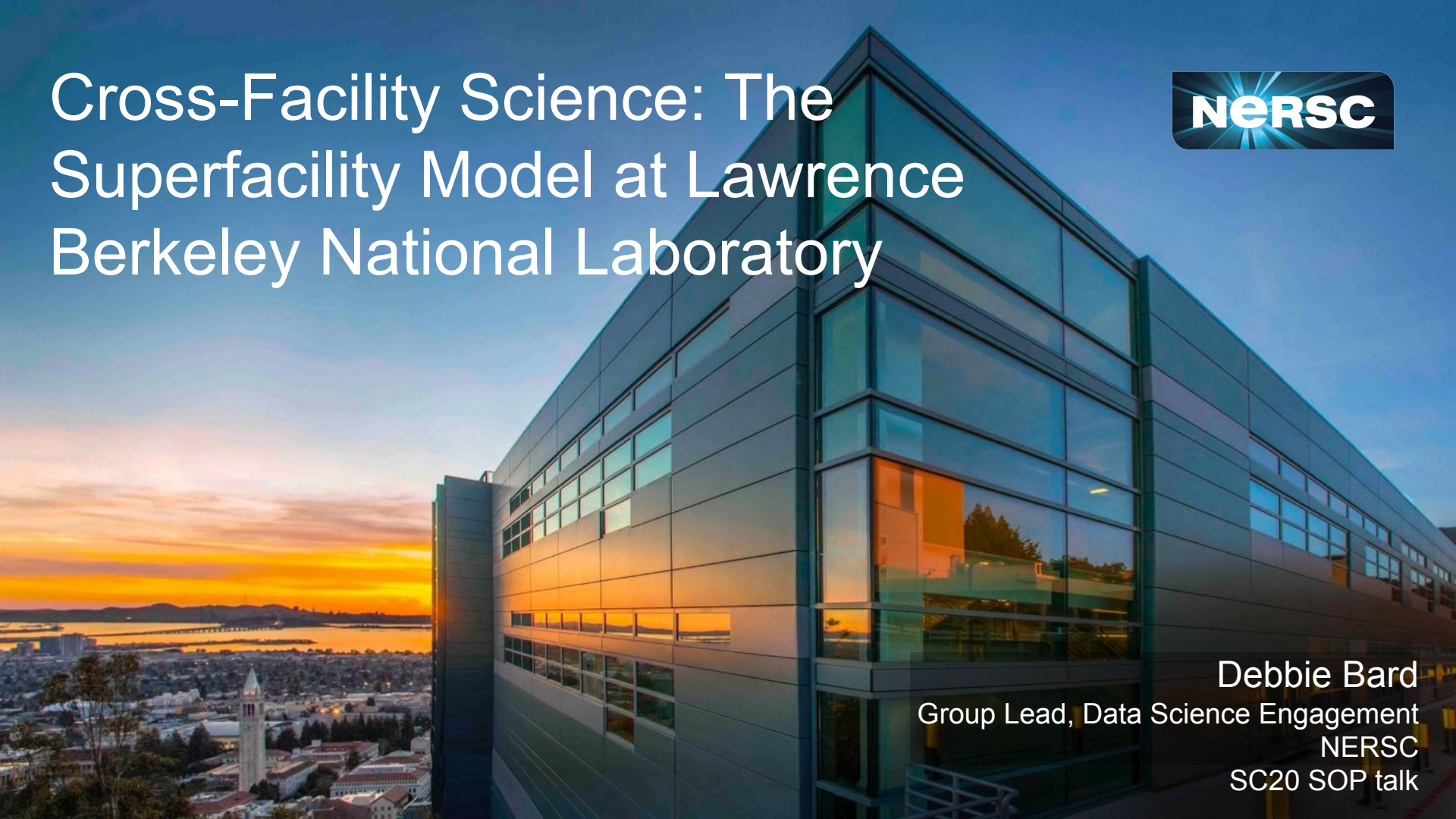
Cross-Facility Science: The Superfacility Model at Lawrence Berkeley National Laboratory

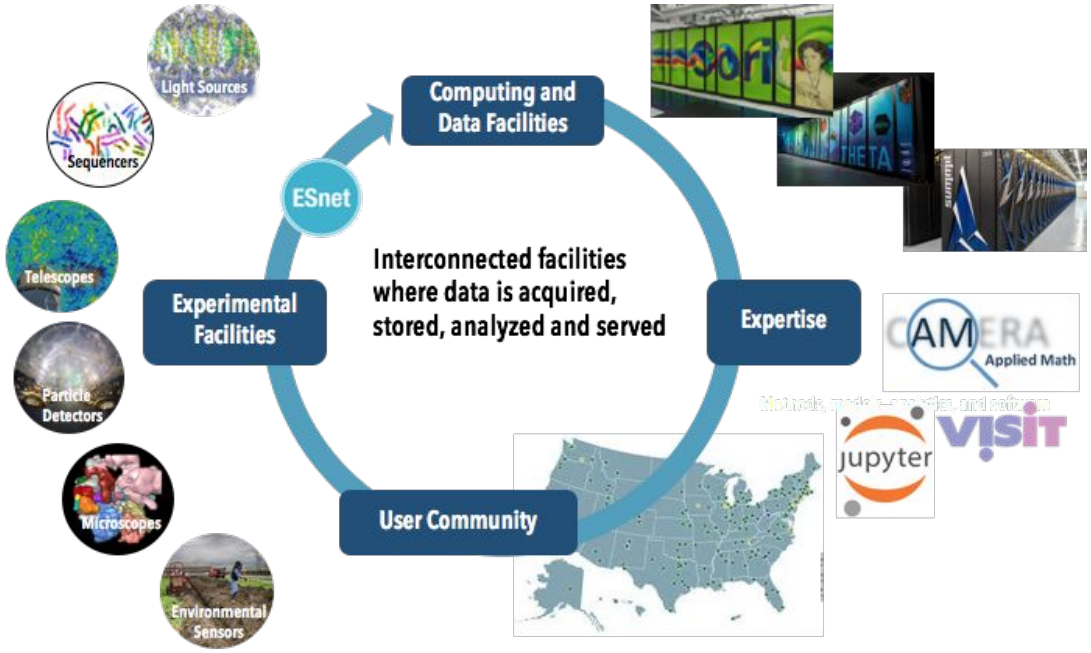


Debbie Bard
Group Lead, Data Science Engagement
NERSC
SC20 SOP talk

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

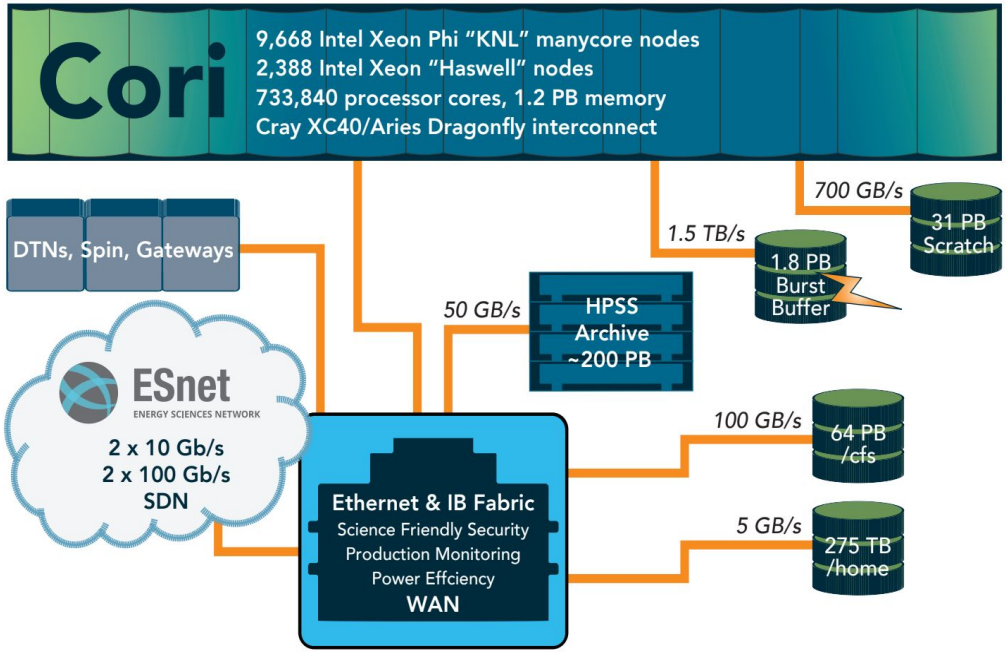
Superfacility@ LBNL: NERSC, ESnet and CRD working together to support experimental science

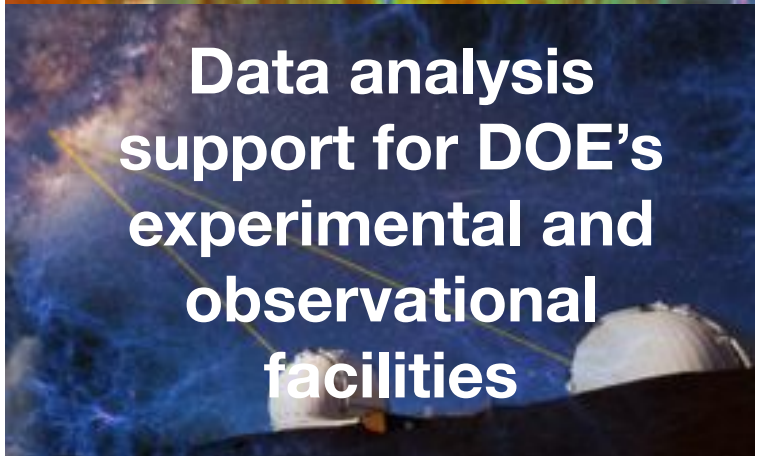
- A model to integrate experimental, computational and networking facilities for reproducible science
- Enabling new discoveries by coupling experimental science with large scale data analysis and simulations



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

7,000+ Users, 800+ Projects
2000+ NERSC citations per year





What features do these experiments need from HPC facilities?

More compute hours

Large volumes of data storage

What features do these experiments need from HPC facilities?

More compute hours

Scalable IO libraries

Large volumes of data storage

Scalable analytics codes

IO patterns with small/random reads/writes

What features do these experiments need from HPC facilities?

More compute hours

Scalable IO libraries

Large volumes of data storage

Scalable analytics codes

Deadline computing

IO patterns with small/random reads/writes

Interactive access and Jupyter

Co-scheduling compute with experiments

What features do these experiments need from HPC facilities?

More compute hours

Scalable IO libraries

Large volumes of data storage

Scalable analytics codes

Deadline computing

IO patterns with small/random reads/writes

Interactive access and Jupyter

Sharing data via web portals:
1. real-time feedback of analysis results
2. access to archived data

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

Co-scheduling compute with experiments

What features do these experiments need from HPC facilities?

More compute hours

Scalable IO libraries

Large volumes of data storage

Resilient workflows to run across multiple computing resources

Scalable analytics codes

Deadline computing

IO patterns with small/random reads/writes

Interactive access and Jupyter

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

Dedicated resources for pipeline/workflow management

Sharing data via web portals:
1. real-time feedback of analysis results
2. access to archived data

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

Co-scheduling compute with experiments

What features do these experiments need from HPC facilities?

More compute hours

Scalable IO libraries

Large volumes of data storage

Resilient workflows to run across multiple computing resources

Scalable analytics codes

Deadline computing

IO patterns with small/random reads/writes

Interactive access and Jupyter

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

Dedicated resources for pipeline/workflow management

Sharing data via web portals:

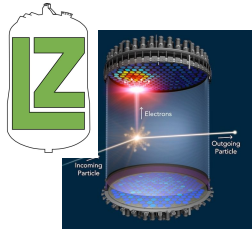
1. real-time feedback of analysis results
2. access to archived data

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

Co-scheduling compute with experiments

Automate everything!

The LBNL Superfacility 'project' coordinates our work 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

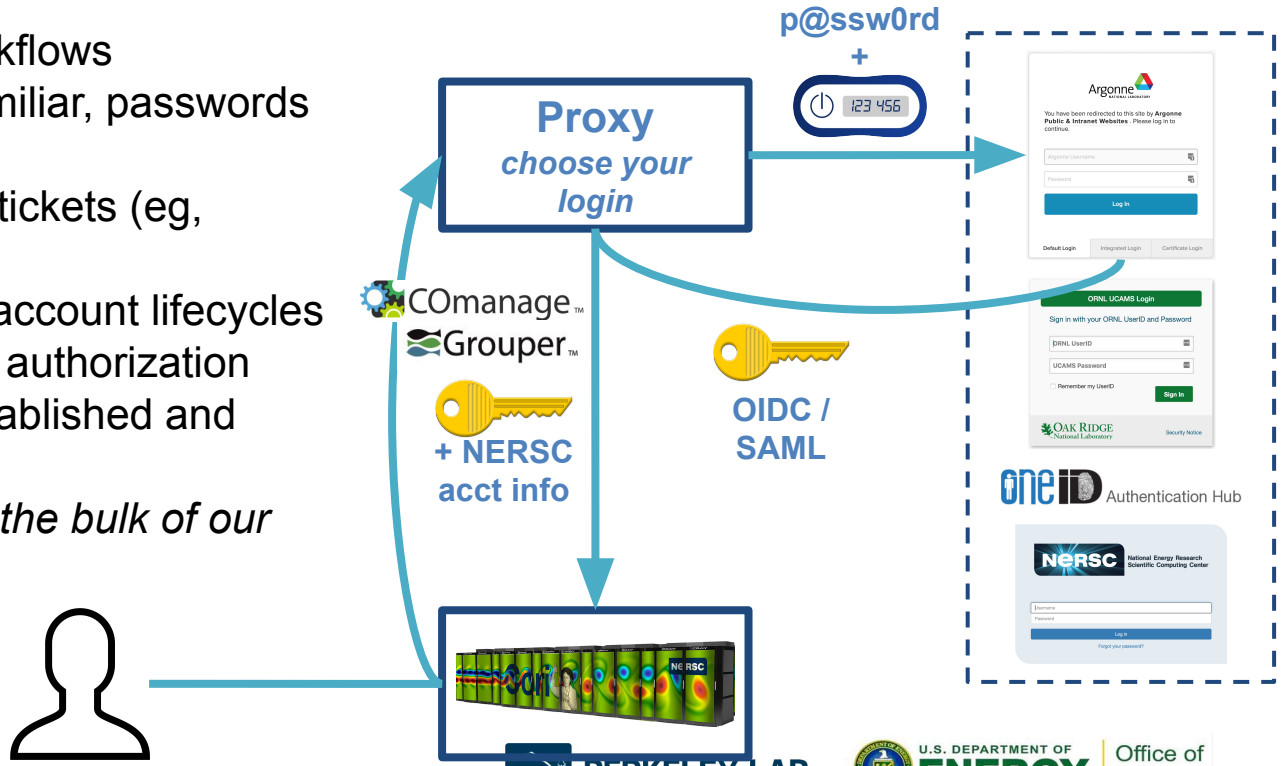
LCLS

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.

Federated Identity (FedID) allows a person to use a single digital identity across multiple organizations

- Simplifies cross-facility workflows
- Users have fewer, more familiar, passwords and login pages
- NERSC has fewer support tickets (eg, password resets)
- Home institution manages account lifecycles
- NERSC still manages local authorization
- Core technology is well-established and mature
- *Policy/trust decisions were the bulk of our analysis*



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*

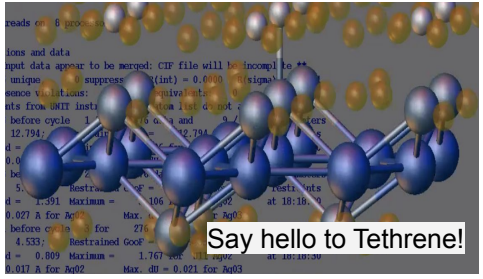
docker RANCHER kubernetes

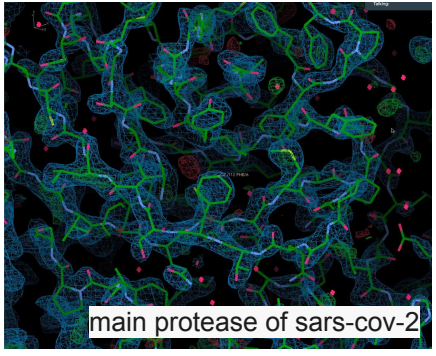
Some projects using Spin:

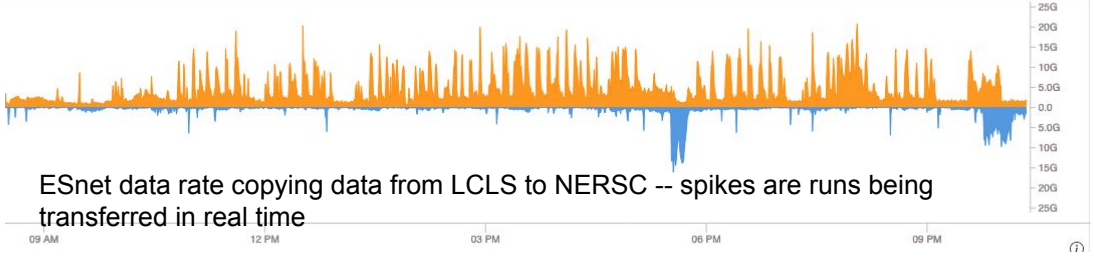
	Track and compare analyses of nightly sky surveys	science gateway
ESS-DIVE	Classify and store reusable earth sciences data	data repository
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY	Manage production genomic workflows and data at scale	science gateway
LZ	Process real-time events for dark matter detection	workflow manager
The Materials Project materialsproject.org	Explore materials properties or build simulated materials	science gateway

LCLS-II is using NERSC for real-time data analysis

- Several experiments at the LCLS-II (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

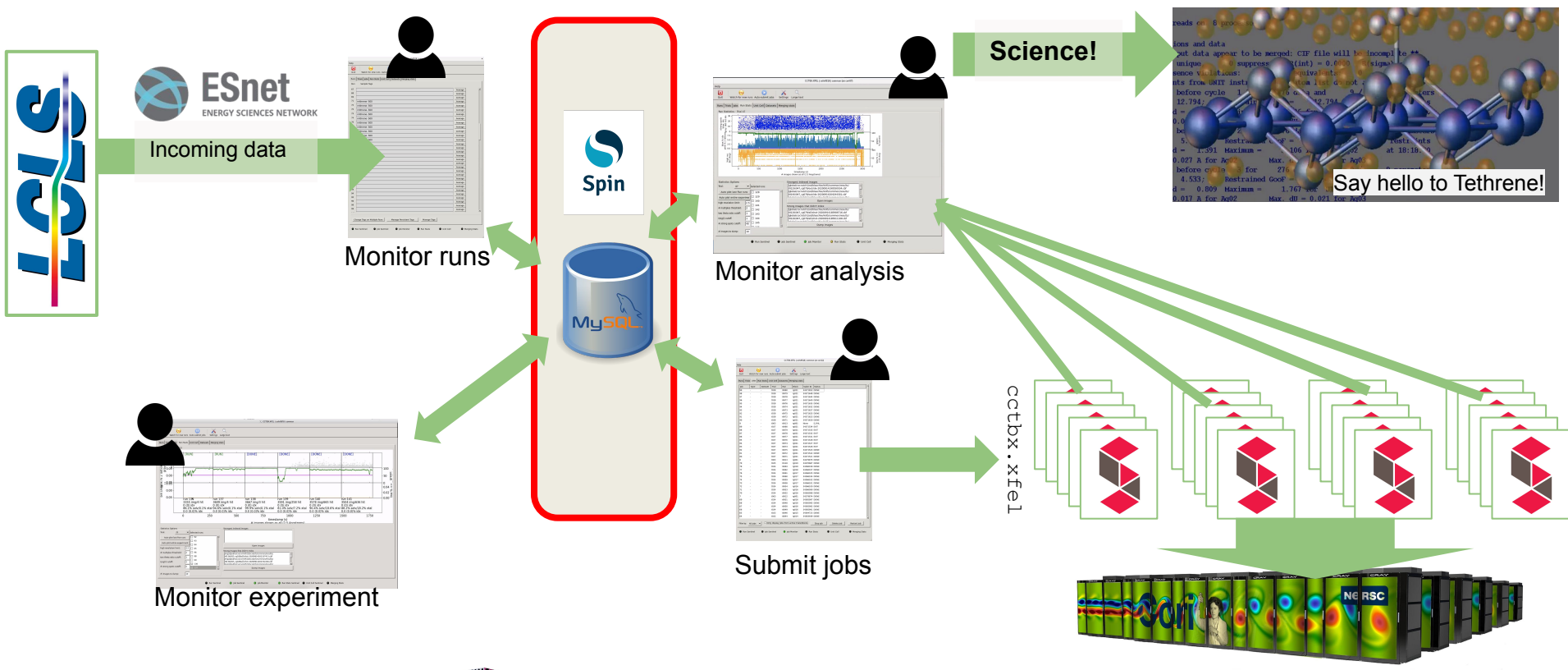

Say hello to Tethrene!


main protease of sars-cov-2



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

Collaborative Distributed Data Analysis with Spin

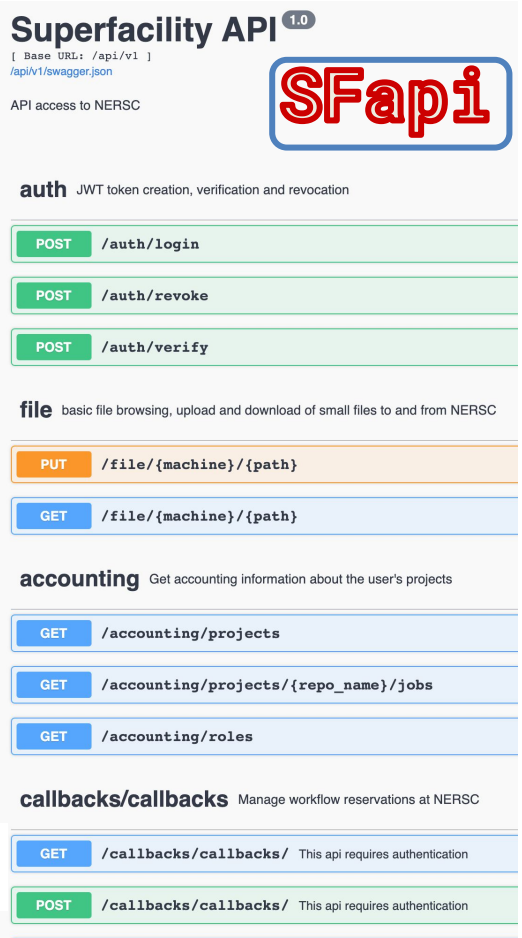


Machine-readable supercomputers: the Superfacility API

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

Endpoints currently prototyped:

`/accounting`	retrieve allocation info for a user or project
`/auth`	obtain OAuth2 authentication tokens (JWTs)
`/callbacks`	register callbacks for asynchronous/chained operations
`/file`	browse, upload, and download files
`/health`	retrieve system health status
`/jobs`	submit jobs and check job status
`/transfer`	move data with Globus or between NERSC storage tiers
`/reservations`	submit and manage future compute reservations



NERSC ESnet ENERGY SCIENCES NETWORK CRD COMPUTATIONAL RESEARCH DIVISION

https://api.nersc.gov/

The Superfacility API: sustainable, scalable automation

- 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'])
```

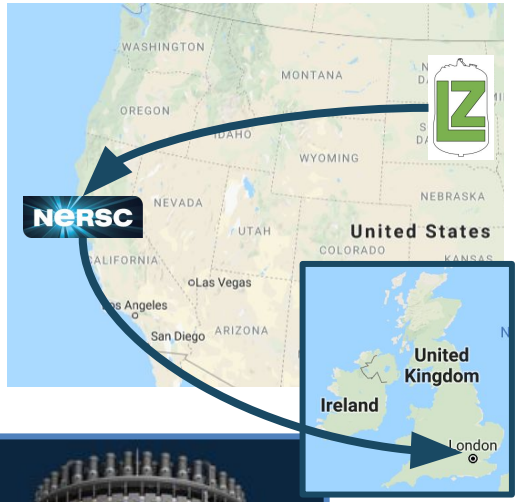
```
Cori is active
```

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 unavailable during this maintenance.', 'status': 'Planned', 'swo': 'true', 'identifier': 'QXg7SbWP3KAeG0mwkyQS', 'updatedate': None}]
```

LZ Dark Matter Detection: watching for DM particles 24/7

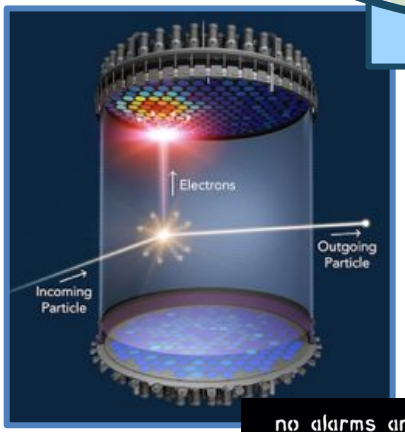


“Simple” online workflow

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

Offline: simulation production and detailed analysis

- **API** to plan for scheduled maintenance
- **API** to avoid manually reading system status from webpage
- **Spin** to share data with many remote users
- **Real-time** access to HPC to monitor running experiment
- **Scalable** simulation code running on GPUs
- **Jupyter** for interactive data analysis at scale

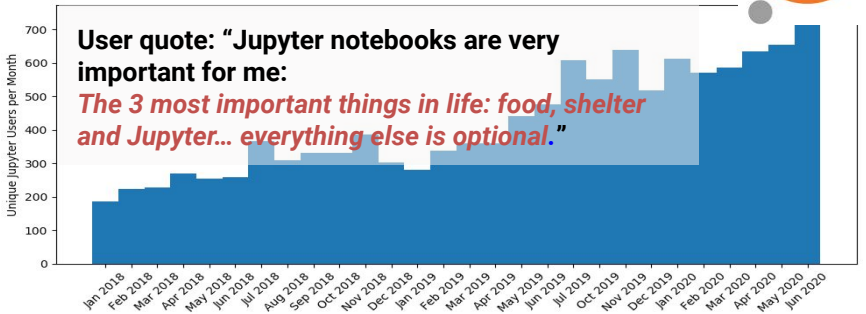


no alarms and no surprises please

Jupyter: supercharge interactive supercomputing

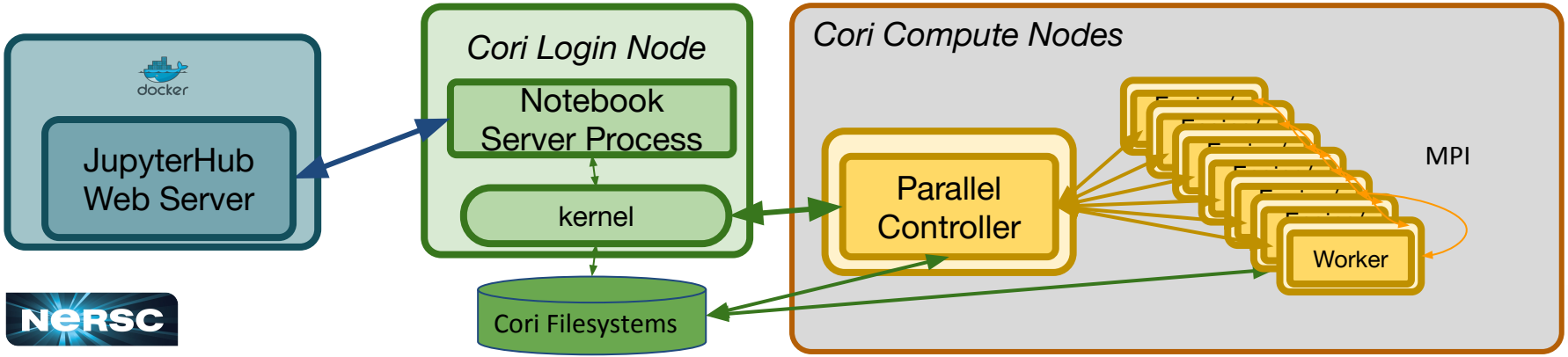
We have deployed an HPC-aware Jupyter service:

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

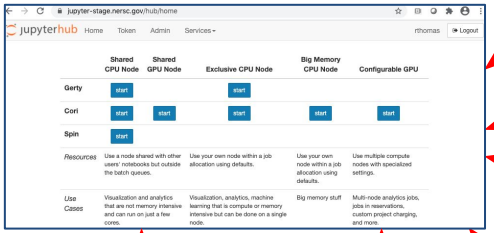


Interactive supercomputing: Jupyter Notebook + HPC Workers

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



Our Hub Leverages NERSC Service APIs



Announcement | NBViewer

Hub Services

Who are you? → sshproxy

Are you a staff user?
What kinds of jobs can you run?
What accounts can you charge to? → Iris

What Shifter images can you run?
Which do you want to run with Jupyter? → SHIFTER

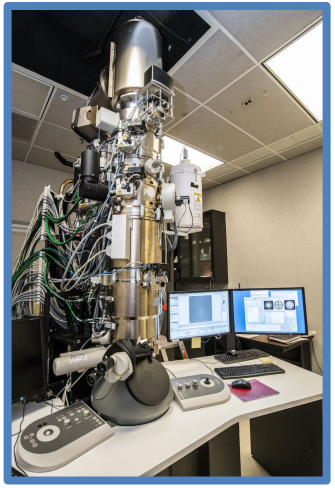
Do you have access to a reservation?
Is the reservation active now? → SFapi

Microservices
Service-oriented architecture

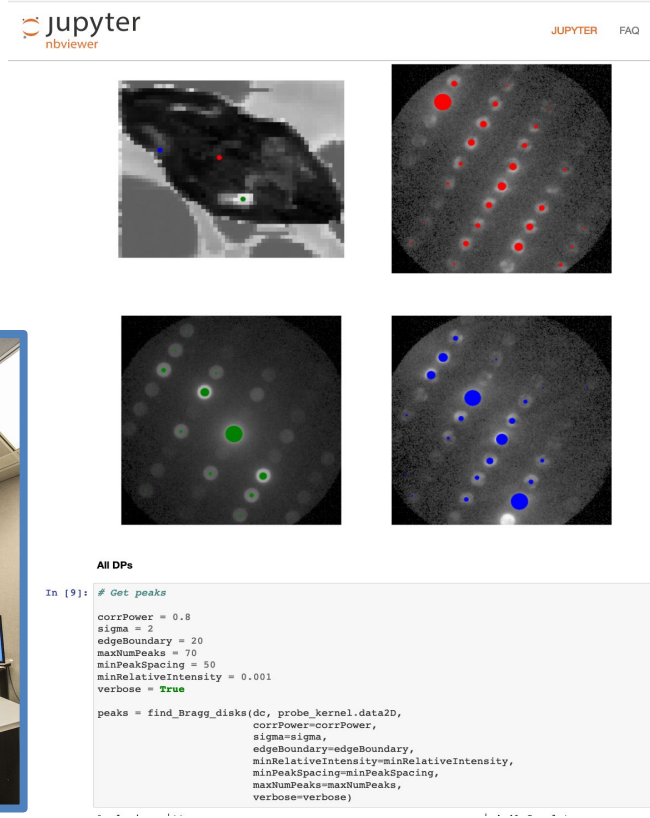
Jupyter and Dask enable interactive exploration and analysis of electron microscope images

MOLECULAR FOUNDRY

- 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
  - No Support for multiple kernels
  - Mismatch leads to cumbersome usage model



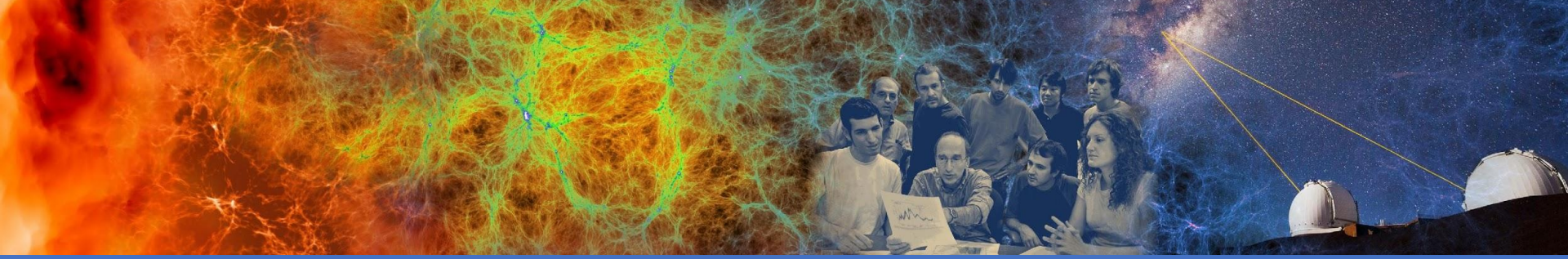
- National Center for Electron Microscopy: Serial processing of 4D image arrays in numpy - Parallelize it!
- **Achieved 20-50x speedup on NCEM Py4DSTEM Notebooks**



Experimental science increasingly needs access to HPC: not just compute+storage, but a whole ecosystem of services

- Automation, timely access to resources, data management and interactivity are key issues
- The LBNL Superfacility project is addressing many of the technical and research needs
  - an **integrated** program of research and technical development to make these workflows seamless and **scalable** across multiple sites and multiple scientific communities
  - See our recent series of demos for more details: https://www.nersc.gov/research-and-development/superfacility/

The LBNL Superfacility Project Team: Debbie Bard, Cory Snavely, Lisa Gerhardt, Jason Lee, Becci Totzke, Katie Antypas, Bill Arndt, Suren Byna, Ravi Cheema, Shreyas Cholia, Mark Day, Bjoern Enders, Aditi Gaur, Annette Greiner, Taylor Groves, Mariam Kiran, Quincey Koziol, Kelly Rowland, Chris Samuel, Ashwin Selvarajan, Alex Sim, David Skinner, Laurie Stephey, Rollin Thomas and Gabor Torok



Thanks!
Questions?

NERSC
ESnet ENERGY SCIENCES NETWORK
CRD COMPUTATIONAL RESEARCH DIVISION
BERKELEY LAB
U.S. DEPARTMENT OF ENERGY | Office of Science