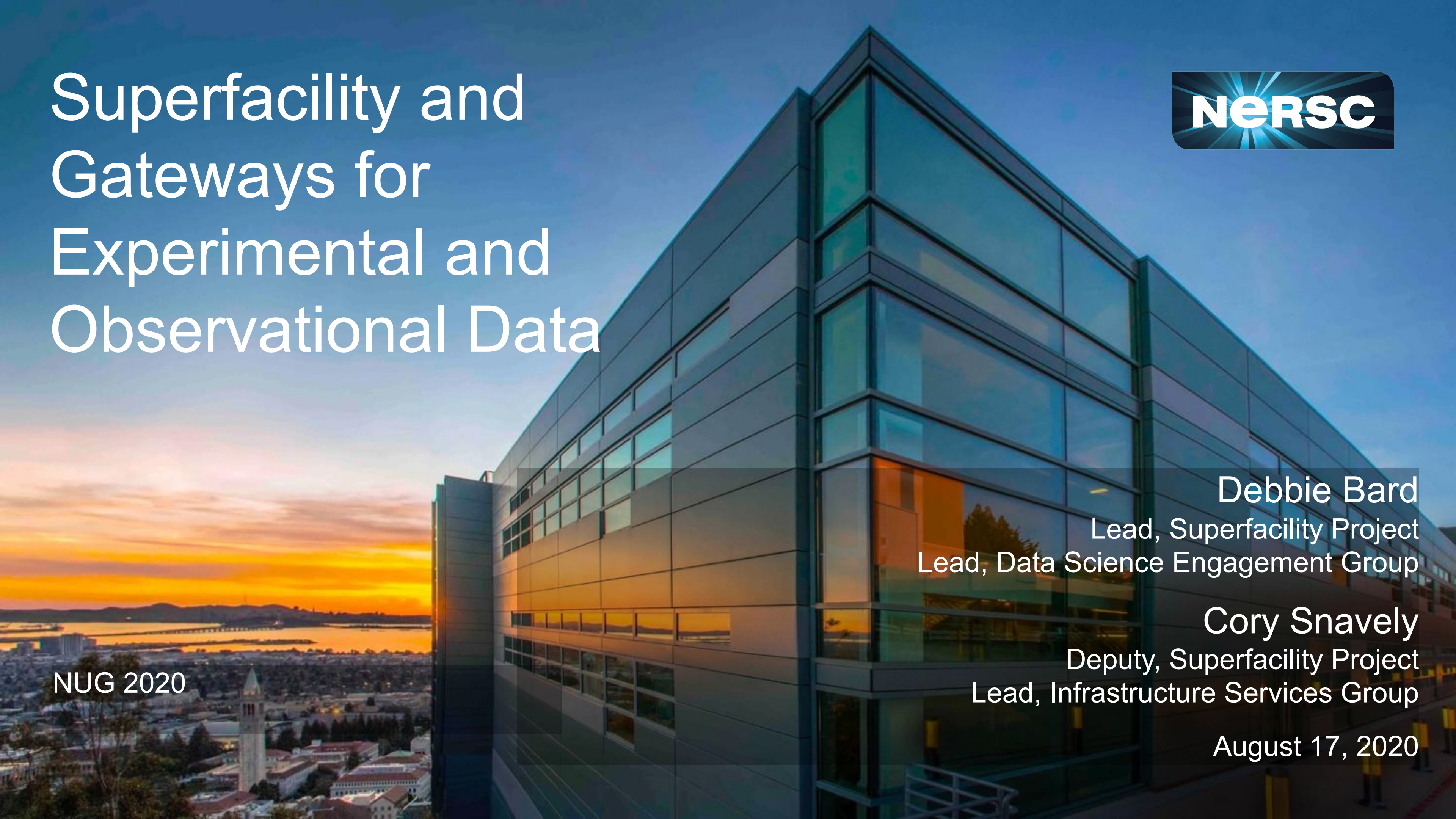


Superfacility and Gateways for Experimental and Observational Data

NUG 2020

Debbie Bard
Lead, Superfacility Project
Lead, Data Science Engagement Group

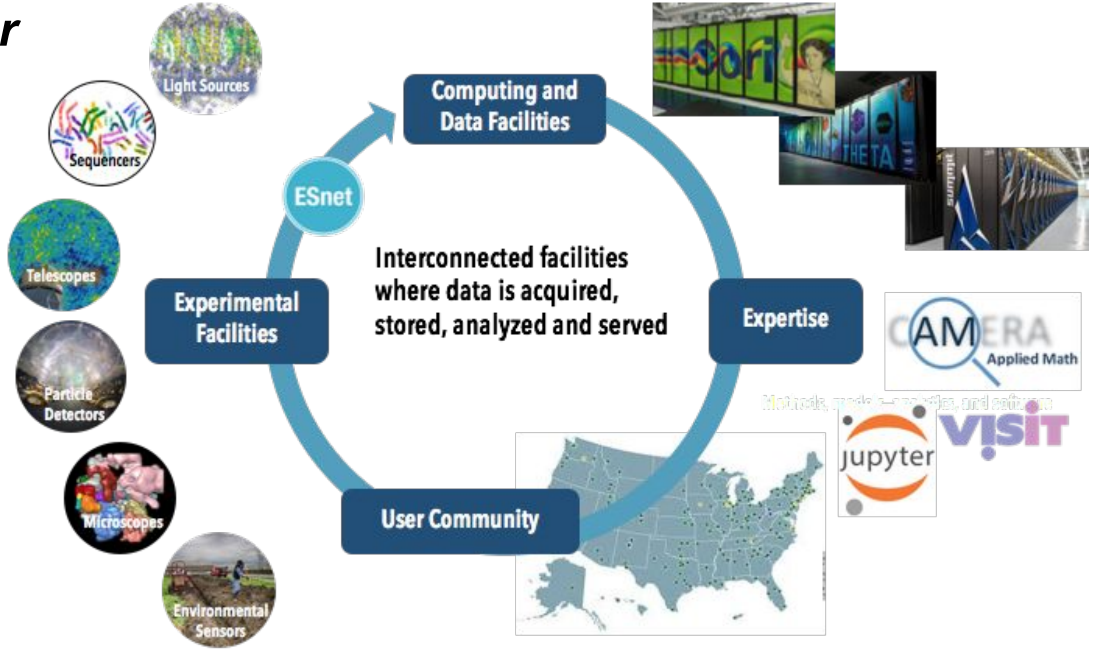
Cory Snavely
Deputy, Superfacility Project
Lead, Infrastructure Services Group

August 17, 2020

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

Superfacility@ LBNL: *NERSC, ESnet and CRD working together*

- 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



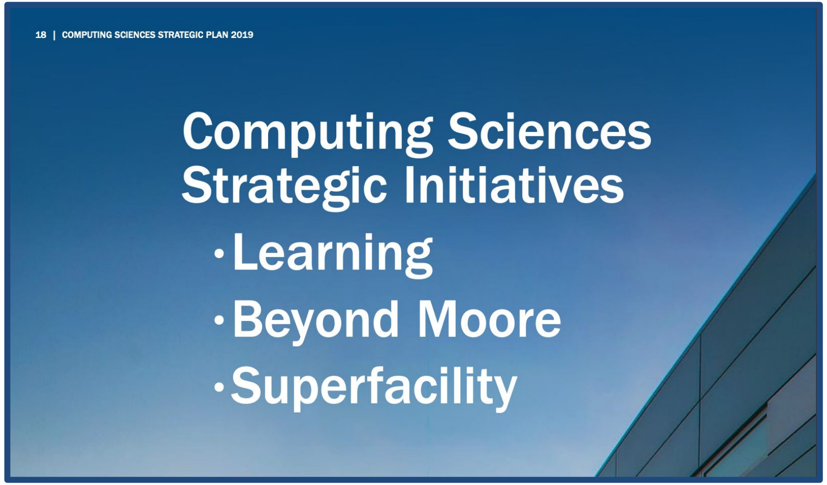
The Superfacility concept is a key part of LBNL strategy to support computing for experimental science

User Engagement

Data Lifecycle

Automated Resource Allocation

Computing at the Edge

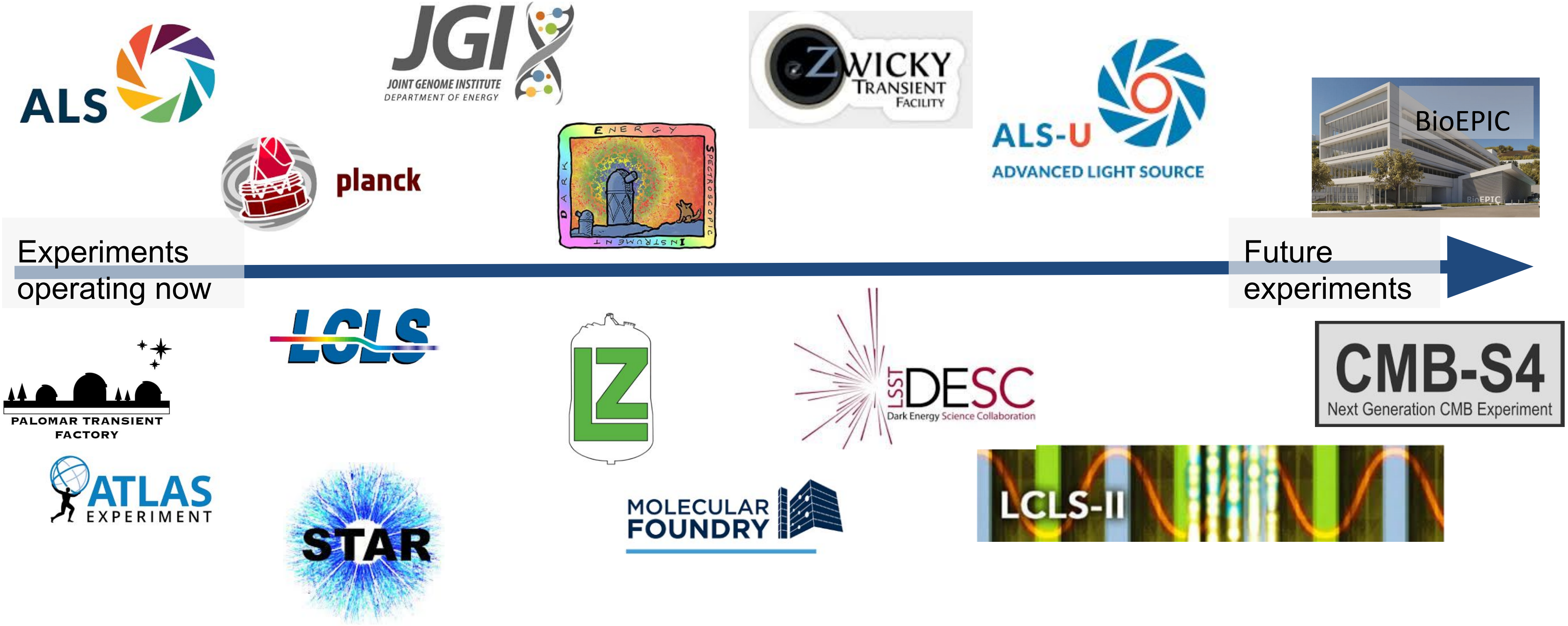


18 | COMPUTING SCIENCES STRATEGIC PLAN 2019

Computing Sciences Strategic Initiatives

- Learning
- Beyond Moore
- Superfacility

NERSC supports many users and projects from DOE SC's experimental and observational facilities

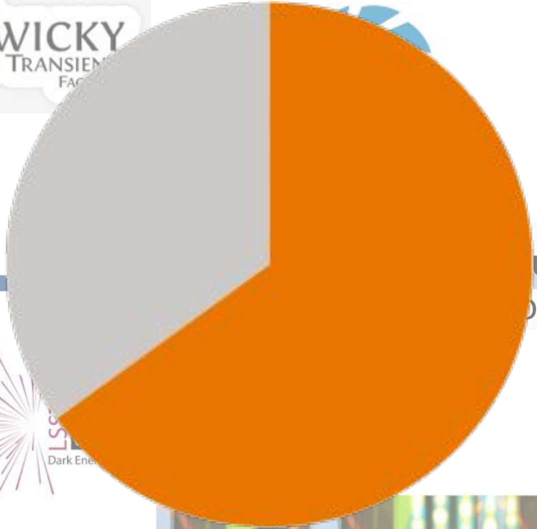


Experiments operating now

Future experiments

NERSC supports many users and projects from DOE SC's experimental and observational facilities

~35% of NERSC projects in 2018 said the primary role of the project is to work with experimental data



ALS

JGI

Zwicky Transient Facility

BioEPIC

Experiments operating now

ure
periments

PALOMAR TRANSIENT FACTORY

LSST Dark Energy

CMB-S4
Next Generation CMB Experiment

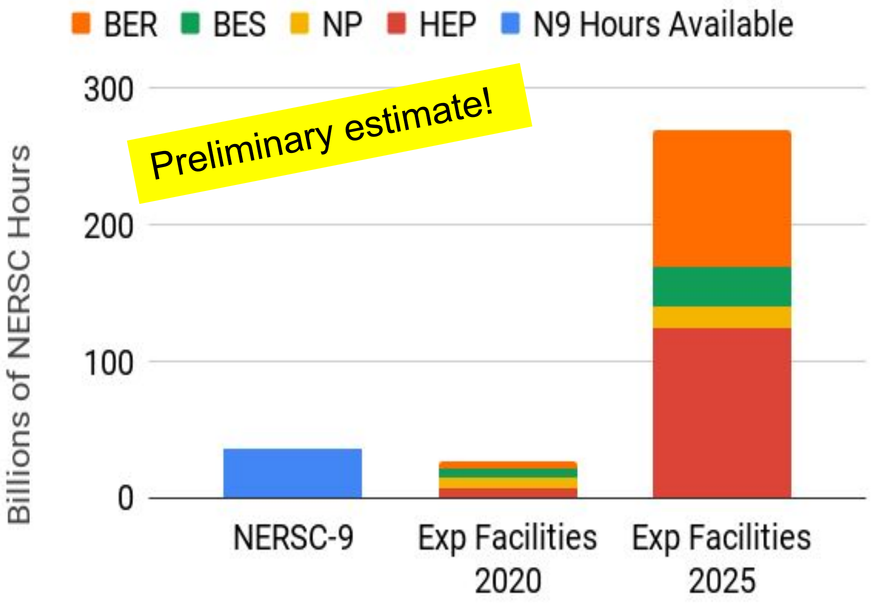
ATLAS EXPERIMENT

STAR

MOLECULAR FOUNDRY

LCLS-II

Compute needs from experimental and observational facilities continues to increase



Taken from Exascale Requirements Reviews

Needs go beyond compute hours:

- High data volumes (today use ~19% of computing hours, but store 78% of data.)
- Real-time (or near) turnaround and interactive access for running experiments
- Resilient workflows to run across multiple compute sites
- Ecosystem of persistent edge services, including workflow managers, visualization, databases, web services…

Compute needs from experimental and observational facilities continues to increase

Taken from Exascale Requirements Reviews

Needs go beyond compute hours:

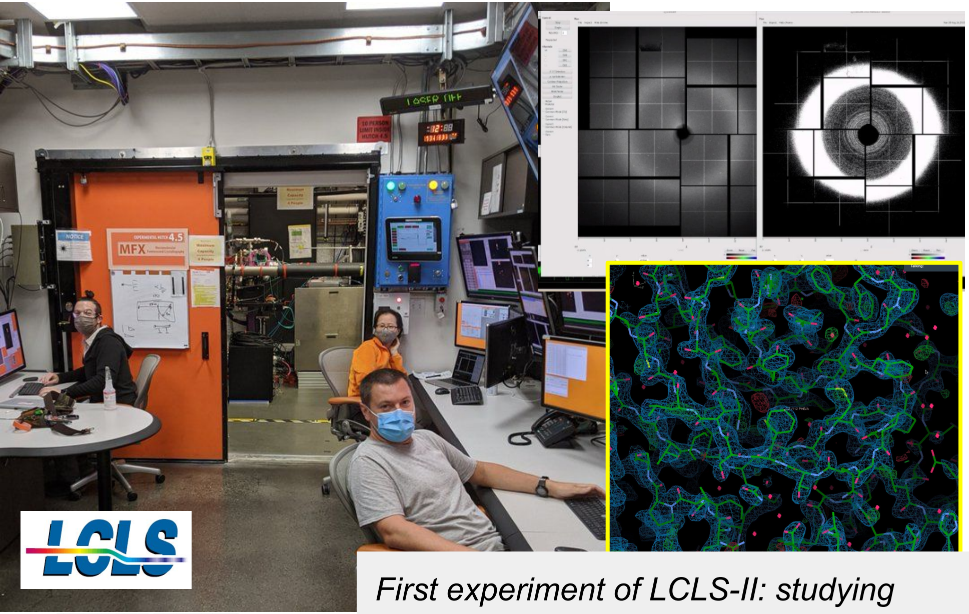
s

including workflow managers, visualization, databases, web services…

You will hear much more about this in the next breakout for the NUGX SIG for Experimental Science Users!

Timing is critical

- Experiments may need HPC feedback: *real-time scheduling*



First experiment of LCLS-II: studying protease for SARS-Cov-2 and inhibitors

NERSC

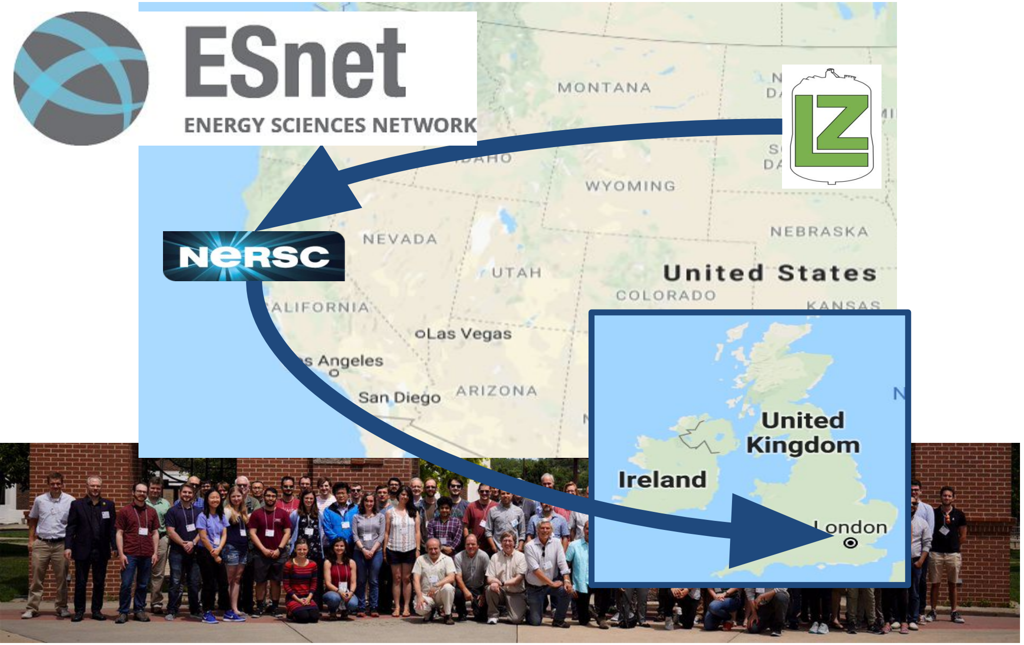
- Workflow may run continuously and automatically: *API access, dedicated workflow nodes*

BERKELEY LAB
Bringing Science Solutions to the World

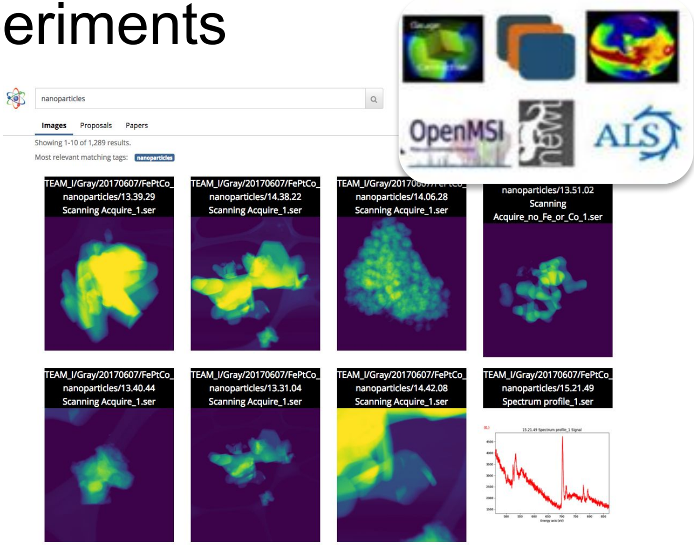
U.S. DEPARTMENT OF ENERGY | Office of Science

Data management is critical

- Experiments move & manage data across sites and collaborators



- Scientists need to search, collate and reuse data across sites and experiments



nanoparticles

Images Proposals Papers

Showing 1-10 of 1,289 results.

Most relevant matching tags: nanoparticles

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.39.29 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/14.38.22 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/14.06.28 Scanning Acquire_1.ser

nanoparticles/13.51.02 Scanning Acquire_no_Fe_or_Co_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.40.44 Scanning Acquire_1.ser

TEAM_I/Gray/20170607/FePtCo_nanoparticles/13.31.04 Scanning Acquire_1.ser

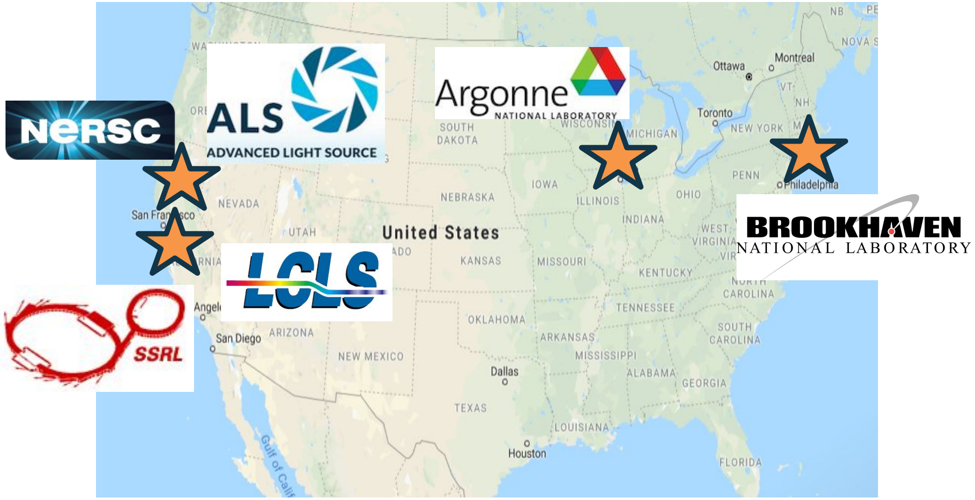
TEAM_I/Gray/20170607/FePtCo_nanoparticles/14.42.08 Scanning Acquire_1.ser

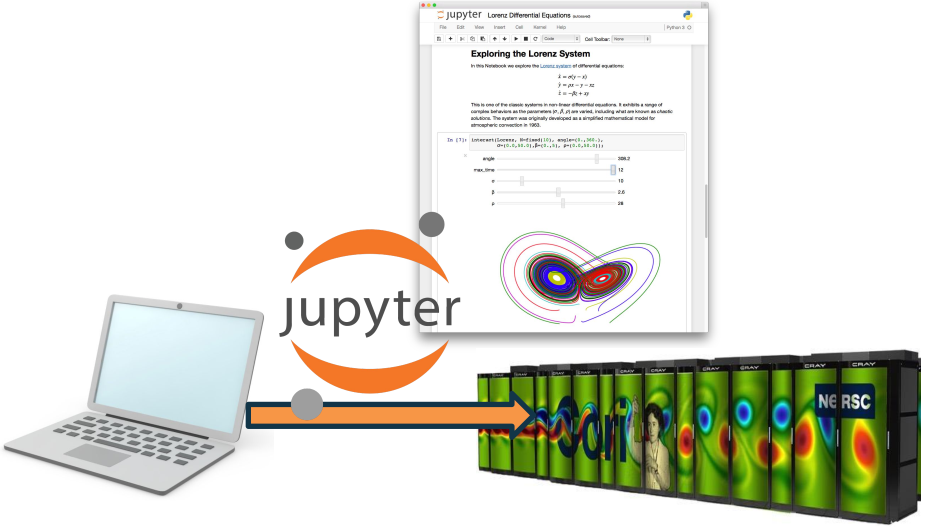
TEAM_I/Gray/20170607/FePtCo_nanoparticles/15.21.49 Spectrum profile_1.ser

Access is critical

- Experiments have their own user communities and policies: Federated ID

- Scientists need access beyond the command line: Jupyter, API…





Lorenz Differential Equations (autosaved)

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters (σ, β, ρ) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0));
```

angle	308.2
max_time	12
σ	10
β	2.6
ρ	28

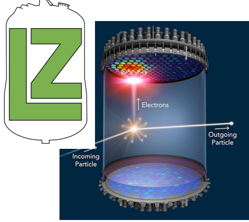
The CS Area Superfacility 'project' coordinates and tracks this work

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
- API-driven automation
- Authentication using Federated Identity

LZ



ALS

FICUS

Facilities Integrating Collaborations for User Science

JGI-EMSL Collaborative Science Call

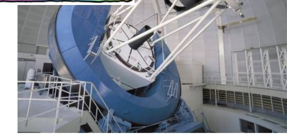
LSST DESC

Dark Energy Science Collaboration

MOLECULAR FOUNDRY

LCLS

DARK ENERGY SPECTROSCOPIC INSTRUMENT



We’ve developed and deployed many new tools and capabilities this year...

Automation *to reduce human effort in complex workflows*

- Released programmable API to query NERSC status, reserve compute, move data etc
- Upgraded Spin: Container-based platform to support workflow & edge services
- Designed federated ID management across facilities

Supported HPC-scale ***Jupyter*** *usage by experiments*

- Scaled out Jupyter notebooks to run on 1000s of nodes
- Developed real-time visualization and interactive widgets
- Curated notebooks, forking & reproducible workflows

Enabled ***time-sensitive workloads***

- Added appropriate scheduling policies, including real-time queues
- Slurm NRE for job pre-emption, advance reservations and dynamic partitions
- Workload introspection to identify spaces for opportunistic scheduling

Deployed ***data management tools*** *for large geographically-distributed collaborations*

- Introduced Globus sharing for collaboration accounts
- Deployed prototype GHI (GPFS-HPSS interface) for easier archiving
- PI dashboard for collaboration management

Superfacility ~~Annual Meeting~~ Demo series

In May/June we held a series of virtual demonstrations of tools and utilities that have been developed to support the needs of experimental scientists at ESnet and NERSC.

- **Recordings available here: https://www.nersc.gov/research-and-development/superfacility/**
  - SENSE: Intelligent Network Services for Science Workflows (*Xi Yang and the SENSE team*)
  - New Data Management Tools and Capabilities (*Lisa Gerhardt and Annette Greiner*)
  - Superfacility API: Automation for Complex Workflows at Scale (*Gabor Torok, Cory Snavely, Bjoern Enders*)
  - Docker Containers and Dark Matter: An Overview Of the Spin Container Platform with Highlights from the LZ Experiment (*Cory Snavely, Quentin Riffard, Tyler Anderson*)
  - Jupyter, Matthew Henderson (*w. Shreyas Cholia and Rollin Thomas*)
- **Planning a second demo series in the Fall as we roll out next round of capabilities**

Priorities for 2020

1. Continue to deploy and integrate new tools, with a focus on the top “asks” from our partner facilities
   - API, Data management tools, Federated ID
2. Resiliency in the PSPS era
   - Working with NERSC facilities team to motivate center resilience
   - Working with experiments to help build more robust workflows
     - eg cross-site data analysis for LZ, DESI, ZTF, LCLS: using ALCC award and LDRD funding
3. Perlmutter prep
   - Key target: at least 4 superfacility science teams can use Perlmutter successfully in the Early Science period

Perlmutter was designed to include features that are good for Superfacility

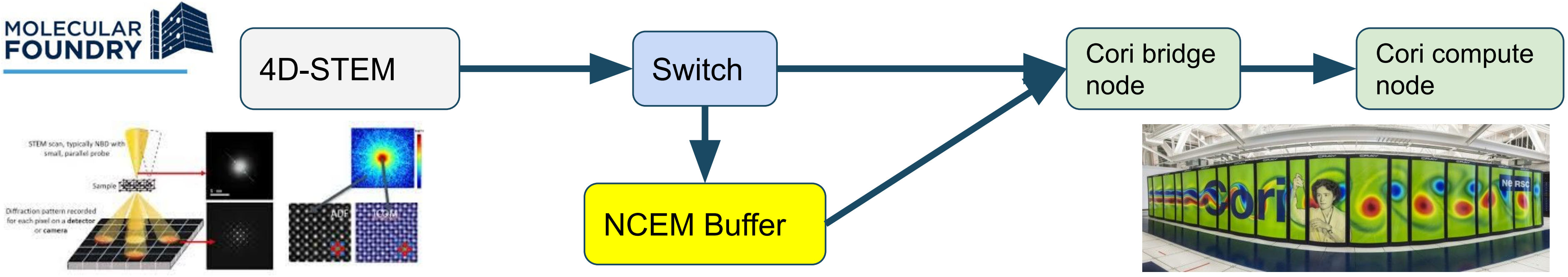
- The Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers combine large scale simulations with experimental data analysis
  - Advocate for and proponent of “team science”

Slingshot Network

- **Slingshot is Ethernet compatible**
  - Blurs the line between the inside/outside machine
  - Allow for seamless external communication
  - Direct interface to storage

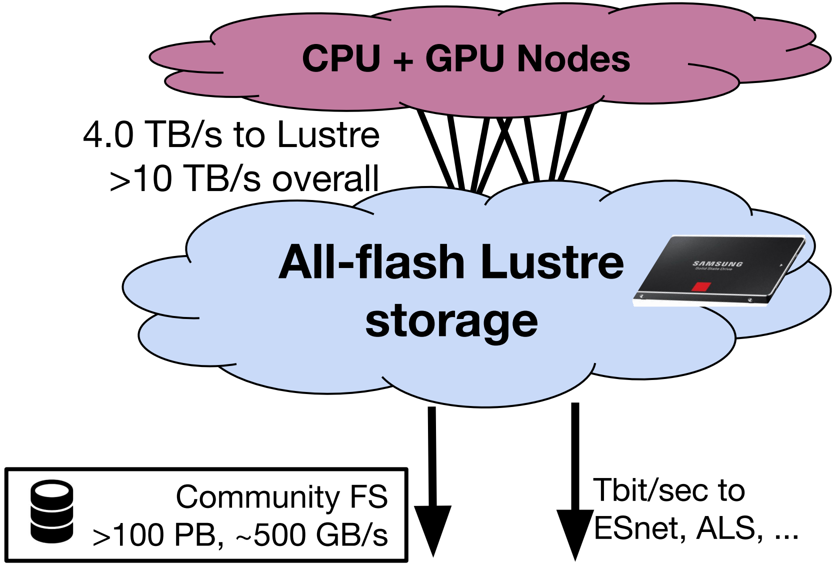
4D-STEM microscope at NCEM will directly benefit from this

- Currently has to use **SDN** and direct connection to NERSC network to stream data to Cori compute nodes
  - uses a buffer into the data flow to send data to Cori via TCP, avoiding packet loss


MOLECULAR FOUNDRY
4D-STEM
Switch
NCEM Buffer
Cori bridge node
Cori compute node
STEM scan, typically NBD with small, parallel probe
Sample
Diffraction pattern recorded for each pixel on a detector or camera
ADF
iCOM
Cori
NERSC

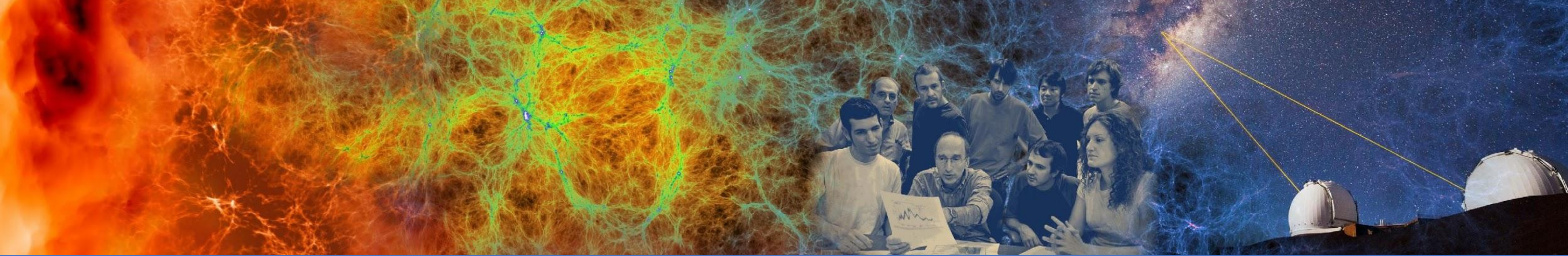
All-Flash scratch Filesystem

- **Fast across many dimensions**
  - 4 TB/s sustained bandwidth
  - 7,000,000 IOPS
  - 3,200,000 file creates/sec
- **Optimized for NERSC data workloads**
  - NEW small-file I/O improvements
  - NEW features for high IOPS, non-sequential I/O



Astronomy (and many other) data analysis workloads will directly benefit from this

- IO-limited pipelines need random reads from large files and databases



Demo: a Science Gateway in 5 Minutes

Spin

Motivation for Spin

“ How can I run services alongside HPC that can…

… access file systems
… access HPC networks
… scale up or out
… use custom software

… outlive jobs (persistence)
… *schedule* jobs / workflows
… stay up when HPC is down
… be available on the web

and are managed by my project team? ”

databases
data archives

workflow
managers

web sites
science gateways

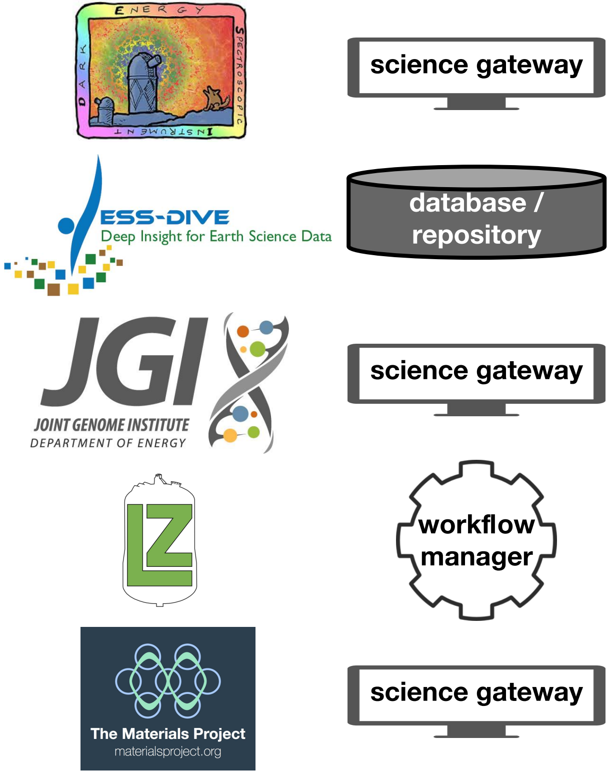
Many Projects Need More Than HPC

Spin answers this need.

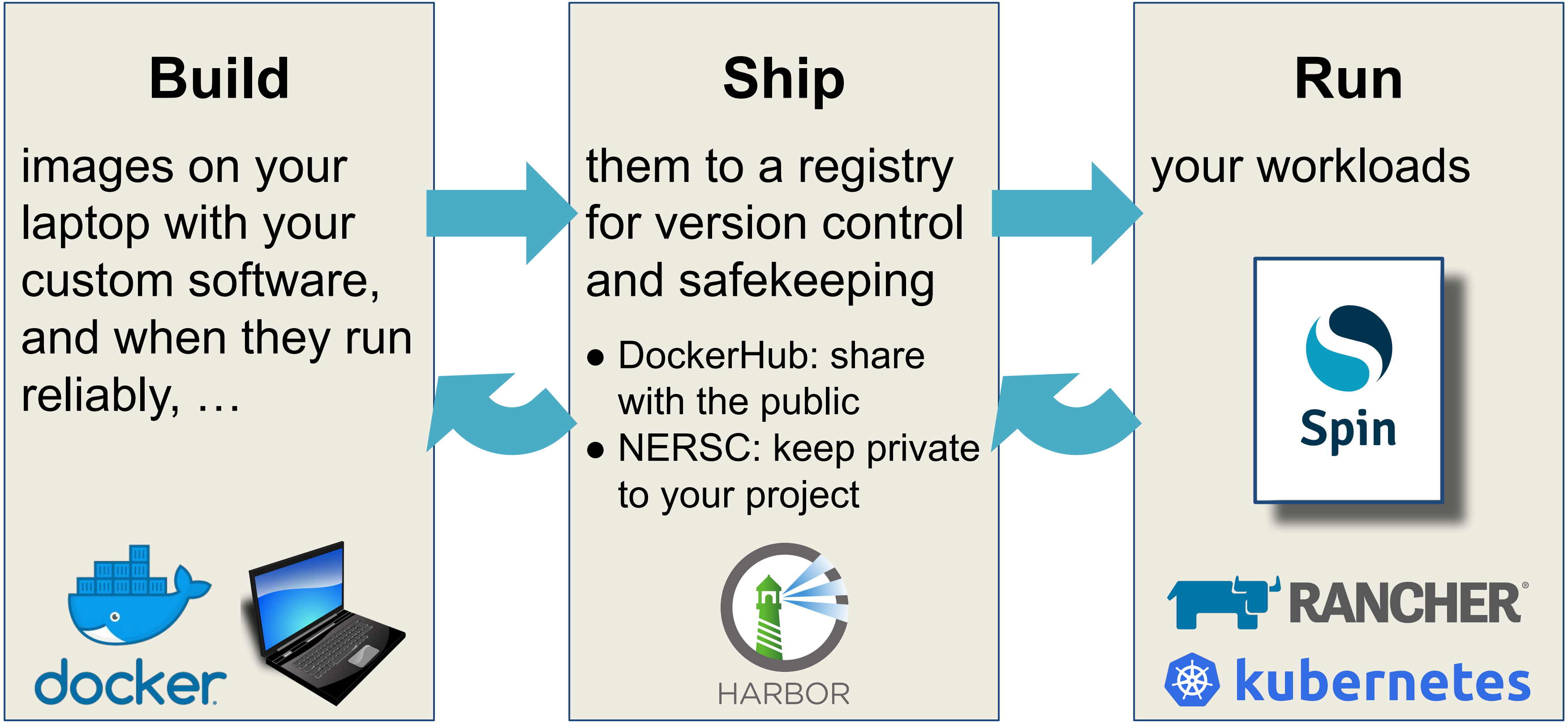
Users can deploy their own **science gateways, workflow managers, databases, and other network services** with Docker containers.

- *Use public or custom software images*
- *Access HPC file systems and networks*
- *Orchestrate complex workflows*
- *...on a secure, scalable, managed platform*

docker RANCHER® kubernetes



Spin Embraces the Docker Methodology



Build

images on your laptop with your custom software, and when they run reliably, …

Ship

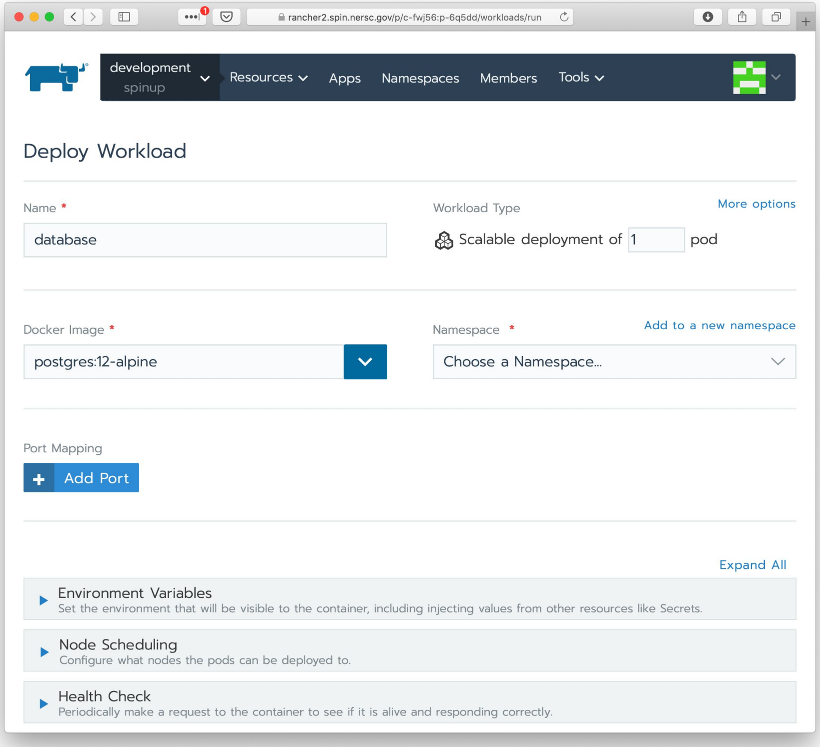
them to a registry for version control and safekeeping

- DockerHub: share with the public
- NERSC: keep private to your project

Run

your workloads

Use a UI, Dockerfile, YAML Declarations…



Deploy Workload

Name * database

Workload Type: Scalable deployment of 1 pod

Docker Image * postgres:12-alpine

Namespace * Choose a Namespace...

Port Mapping

Environment Variables
Set the environment that will be visible to the container, including injecting values from other resources like Secrets.

Node Scheduling
Configure what nodes the pods can be deployed to.

Health Check
Periodically make a request to the container to see if it is alive and responding correctly.

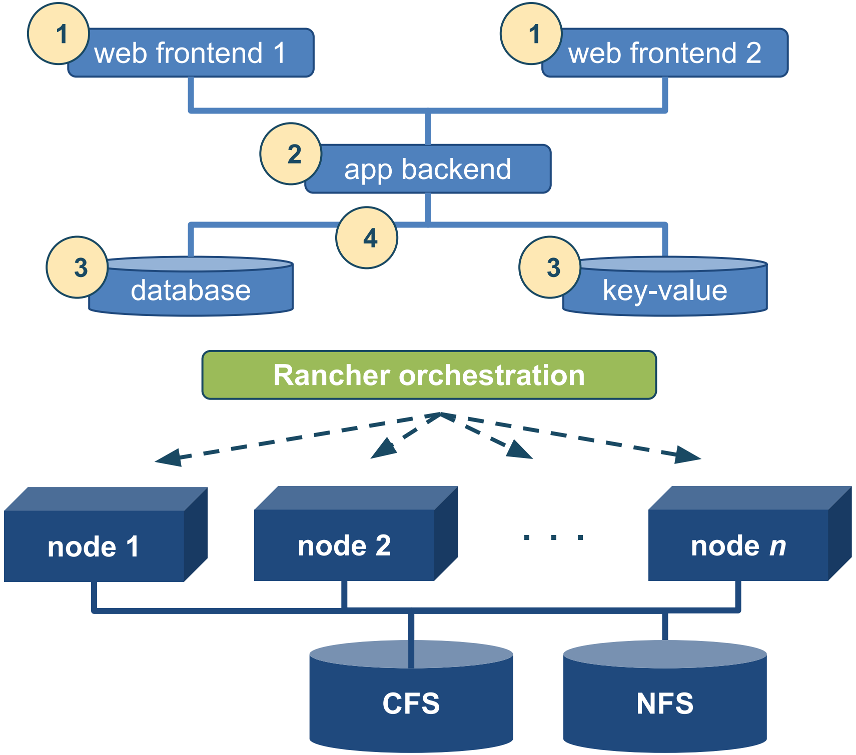
my-project.yml

```
baseType: workload
containers:
  name: ap
  image: f
  imagePul
  environm
    TZ: US
  volumeMo
  - mountP
    name:
    type:
    readOn
...
```

Dockerfile

```
FROM ubuntu:18.04
RUN apt-get update  --quiet -y && \
    apt-get install --quiet -y \
    python-flask
WORKDIR /app
COPY app.py /app
ENTRYPOINT ["python"]
CMD ["app.py"]
```

…to create running services.



A typical example:

1. **multiple nginx frontends**
2. **custom Flask backend**
3. **database or key-value store (dedicated, not shared)**

automatically plumbed into a

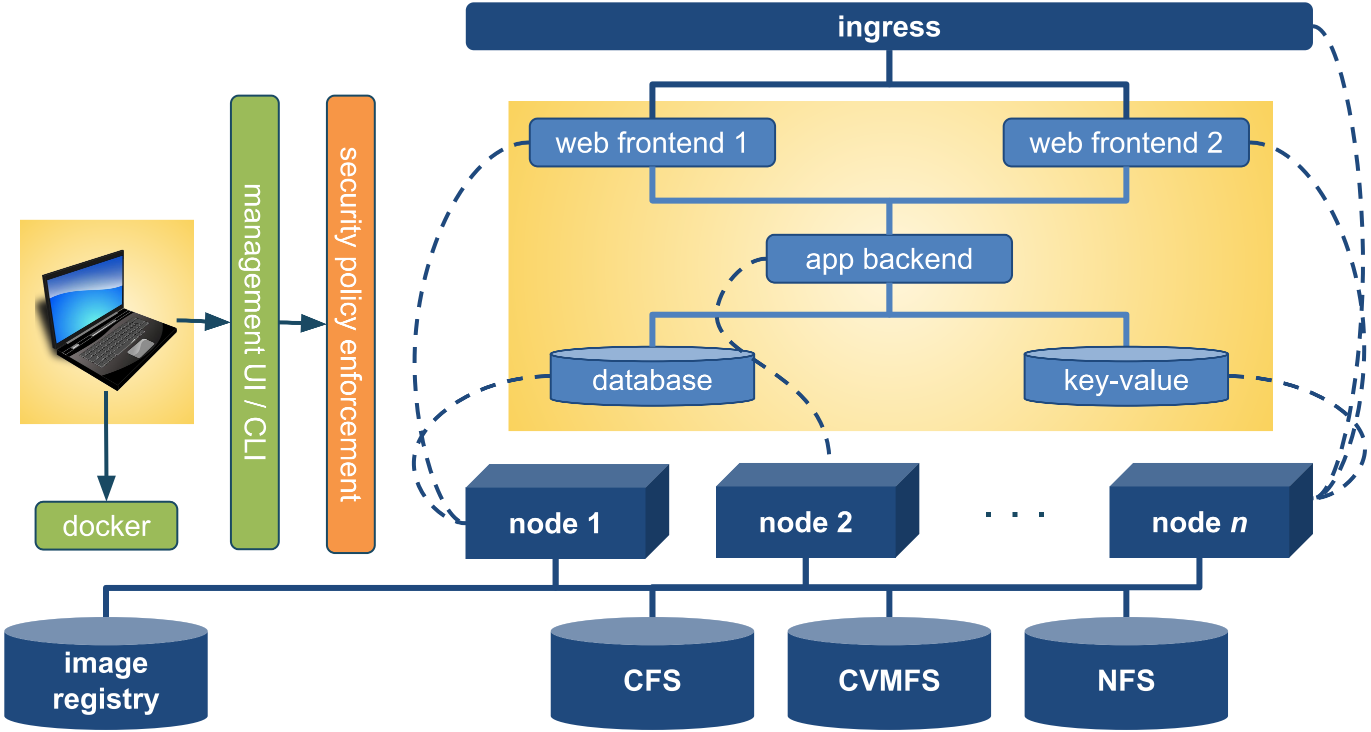
4. **private overlay network.**

Rancher starts all the containers and ensures they stay running.

High-Level Spin Architecture

User-managed

NERSC handles the rest!

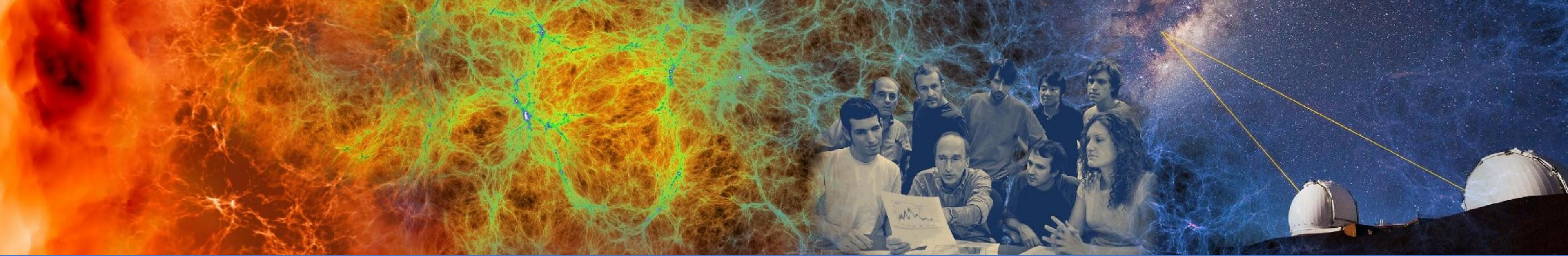


Demo: Creating a Service in Spin

Learn More about Spin

Attend a SpinUp Workshop to learn how you can build your own science gateways!

More info: https://www.nersc.gov/systems/spin/



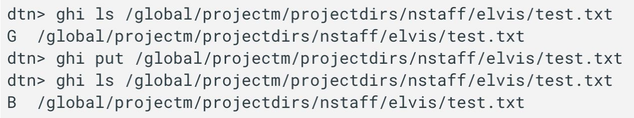
Fin

New API functionality: https://api.nersc.gov/

- Workflow automation needs to interact w/ NERSC w/o a human in the loop:
    - Eg beamline at NSLS-II wants to send data for analysis
    - Requirements based on detailed survey in winter 2019
  - Ask questions like:
    - Is NERSC in maintenance?
    - When are future maintenances scheduled?
    - Is the scratch file system available?
  - Perform actions like:
    - Move my data
    - Launch a job
    - Make a reservation...
- Finalizing authentication model and implementation
  - Not yet visible to users - pending completion and security review
- Staff to contribute via Gitlab-based process

New Data Movement tools deployed

- Large collaborations (eg LZ, LSST-DESC) struggle to manage their data between CFS and HPSS
  - **GHI is deployed to early users**
    - Easy way to archive data from CFS using command line tools
    - Automatically bundles data to optimal HPSS size



```
dtn> ghi ls /global/projectm/projectdirs/nstaff/elvis/test.txt
G  /global/projectm/projectdirs/nstaff/elvis/test.txt
dtn> ghi put /global/projectm/projectdirs/nstaff/elvis/test.txt
dtn> ghi ls /global/projectm/projectdirs/nstaff/elvis/test.txt
B  /global/projectm/projectdirs/nstaff/elvis/test.txt
```

- Experiments often share the data management between multiple staff - we use collab accounts to enable this
  - **Collaboration accounts enabled for Globus sharing**
  - Dedicated endpoint allows specified users to transfer data in as collab user, no extra step needed to manage permissions

- PIs of large teams often have to ask NERSC to chown/chgrp collaboration data when users leave or mess up their permissions
  - **PI Data Dashboard enables these actions via a click of a button**

Areas of Technical Work

Advanced Scheduling; Resiliency

Support forecasted real-time computing demands

Software-Defined Networks; SENSE; Self-Managed Systems

Provide on-demand connectivity, QoS, fault handling, etc

Data Movement; Data Dashboard; HDF5

Simplify data management tasks and optimize data production and analysis

Spin: Containers-as-a-Service Platform

Support “edge services” adjacent to HPC for workflows

API and Federated Identity

Automate it all and use modern cross-facility authentication

Drivers:

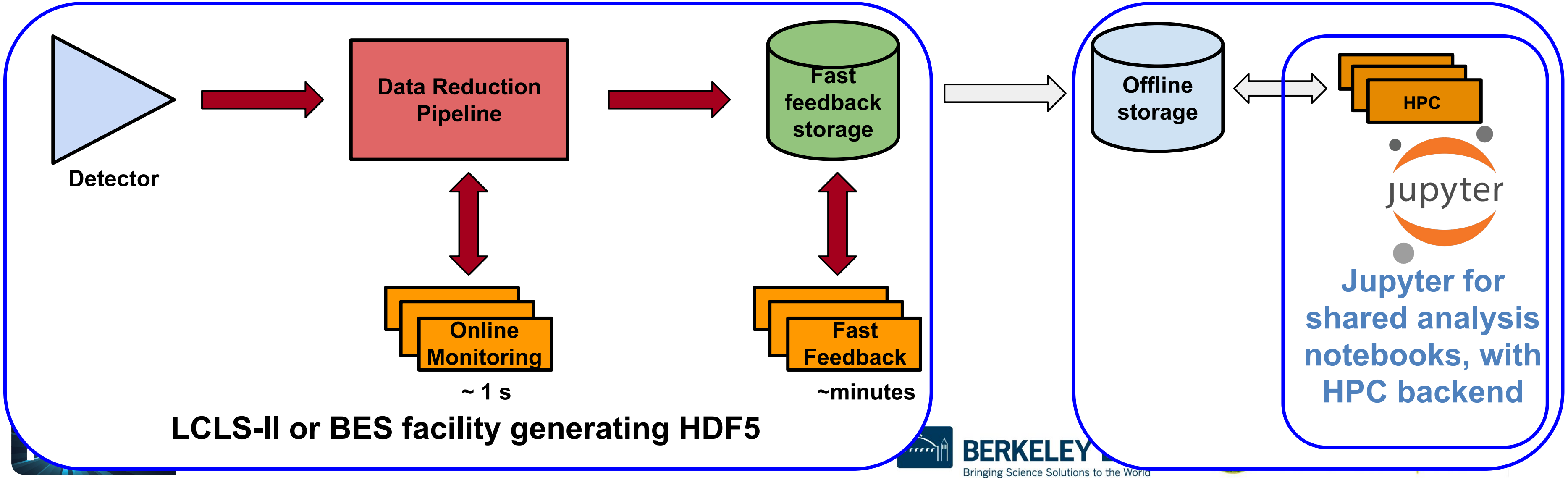
- Complex workflows
- Data-driven projects
- Real-time compute
- Streaming instrument data

LLAna: LCLS-LBNL Data Analytics Collaboration

Pilot to design and deploy a new computing environment for the next generation of free electron lasers: tools for composable workflows, data management and analysis.

HDF5 for high-performance file access and management, designed for LCLS-II needs

Workflow profiling, characterization and optimization for real-time LCLS-II analysis on HPC resources



Jupyter for shared analysis notebooks, with HPC backend

The NERSC-9 Project is Proceeding Well

- Scope, Cost, Schedule ✓
- Facility Upgrade on Track ✓
- Health & Safety Processes ✓
- CD 2/3 ✓
- App Readiness Progress ✓
- Staff Experience ✓
- System Contract Award ✓
- Risks Defined & Managed ✓
- Well Trained CAMS ✓

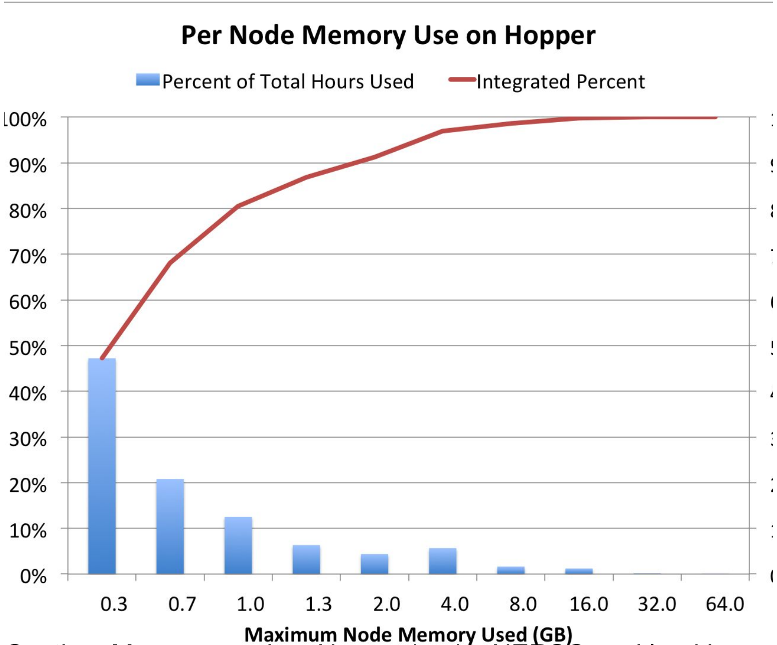
Annual Project Review Nov. 5-6, 2019

Only 1 recommendation: Continue prioritization of hiring a permanent lab project manager

12.5 MVA power upgrade and associated cooling for N9 underway

Hopper Memory Usage

- Feugiat, facilisis mauris.
- Erat arcu lorem donec sceleris
- Parturient

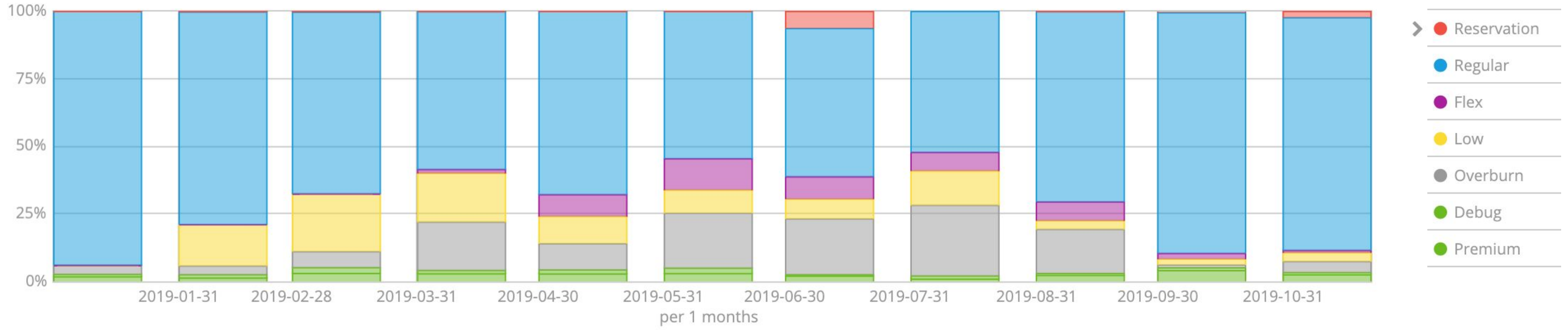


Caption: Memory used on Hopper by the NERSC workload in 2013.

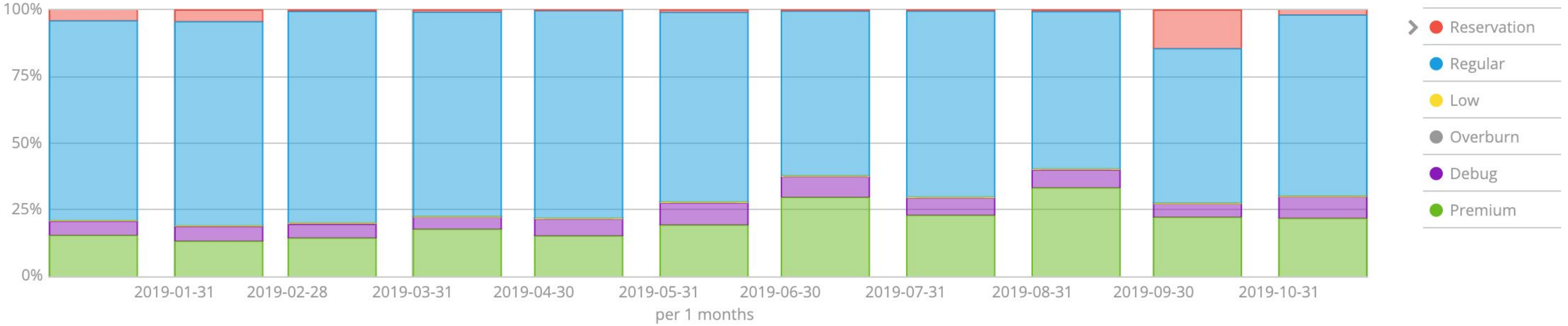
A Table

Column 1	Column 2	Column 3

Cori KNL QOS Usage by Month



Cori Haswell QOS Usage by Month



Alternative Section Divider