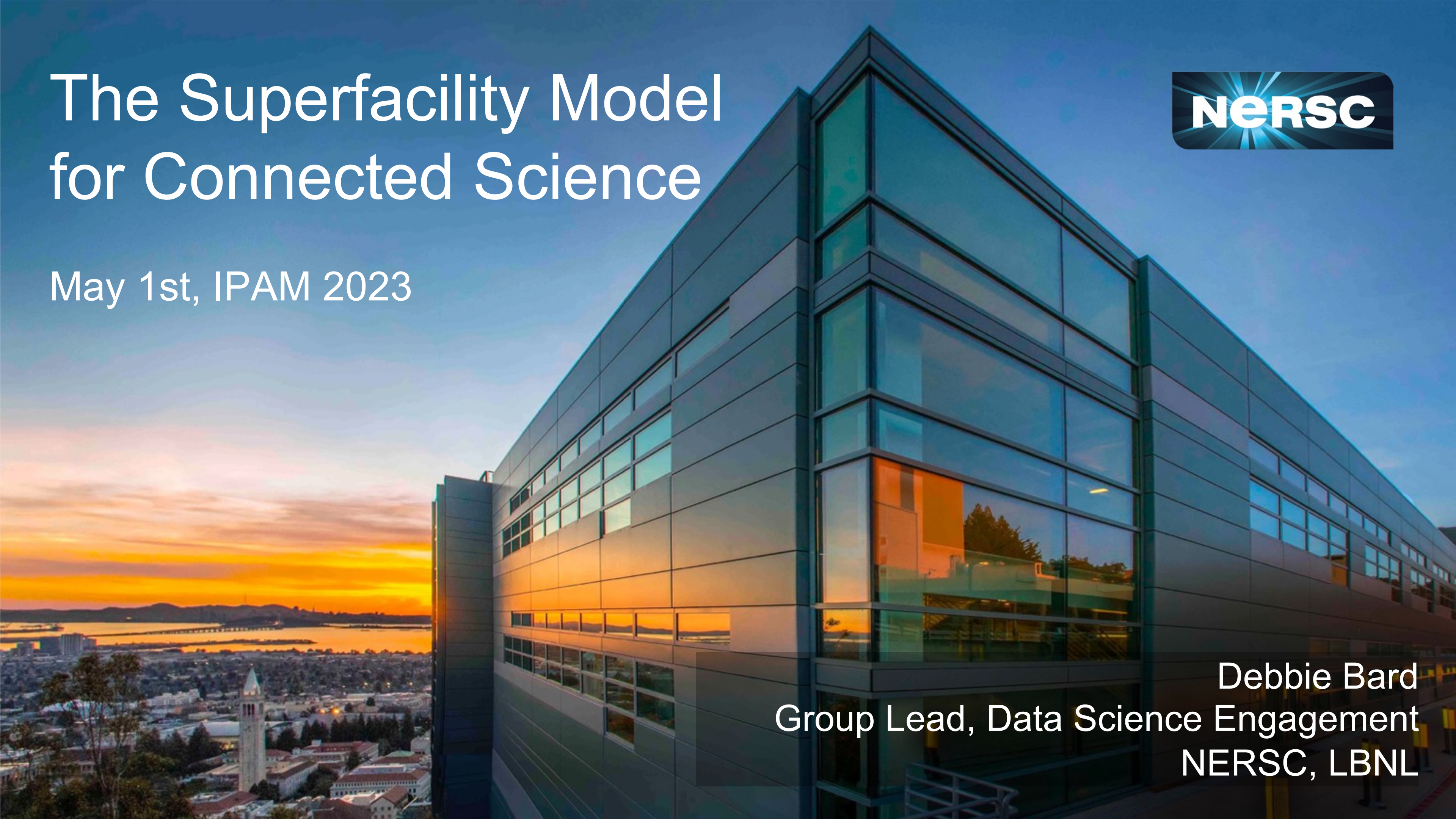


The Superfacility Model for Connected Science

May 1st, IPAM 2023

NERSC

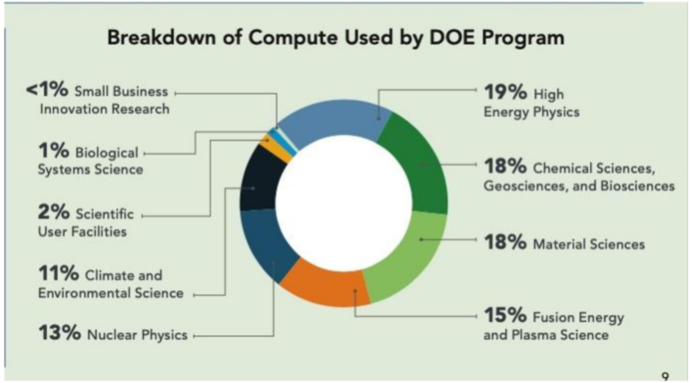
Debbie Bard
Group Lead, Data Science Engagement
NERSC, LBNL

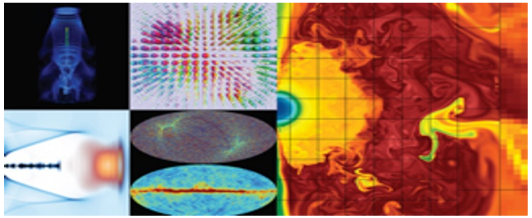
NERSC is the mission High Performance Computing facility for the DOE Office of Science

9,000 Users
1,000 Projects

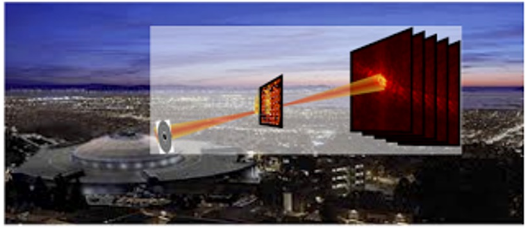
>2,000

Scientific Journal Articles per Year

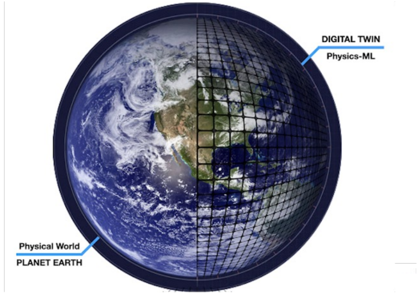




Simulations at scale

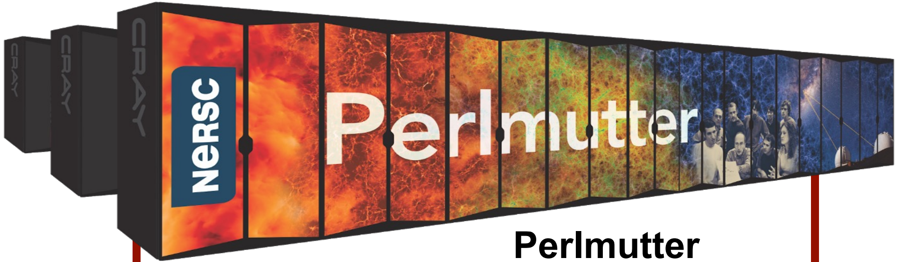


Urgent and interactive computing
Photo Credit: CAMERA



Complex experimental & AI workflows
Photo credit: A depiction of digital twin Earth adapted from the EU's Destination Earth project.

NERSC Center Architecture



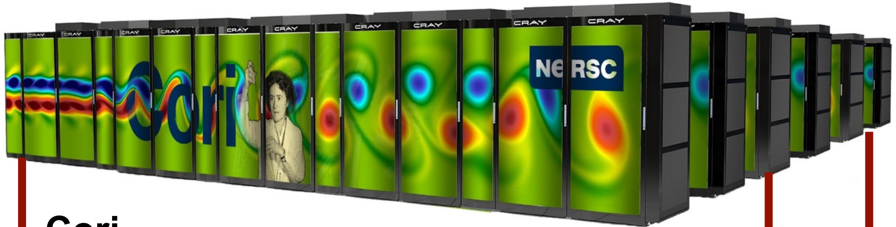
Perlmutter

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

2 PB Burst Buffer

700 GB/s

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

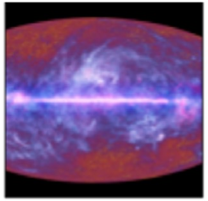
120 PB Community File System

5 GB/s

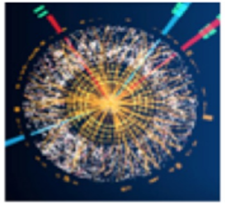
275 TB /home

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

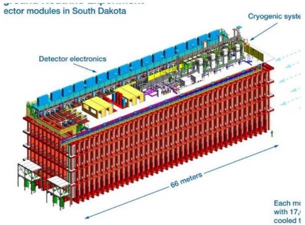

Palomar Transient Factory Supernova


Planck Satellite Cosmic Microwave Background Radiation

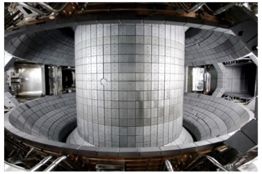
Star Particle Physics


Atlas Large Hadron Collider


APS

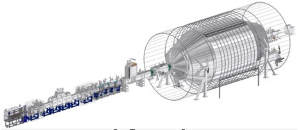


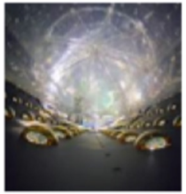
Dune

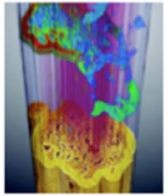

KStar


ARM

GlueX


Katrin

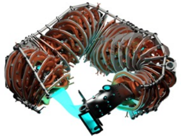

Dayabay Neutrinos


ALS Light Source


LCLS Light Source

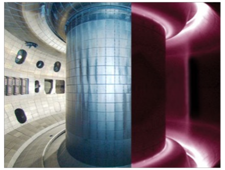

Joint Genome Institute Bioinformatics

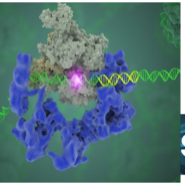

NSLS-II

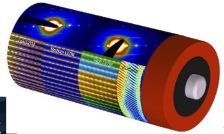

HSX

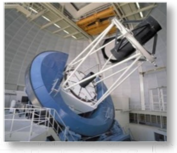

Majorana

AmeriFlux

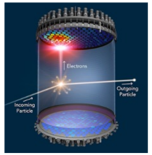

DIII-D


Cryo-EM


NCEM


DESI


LSST-DESC


LZ

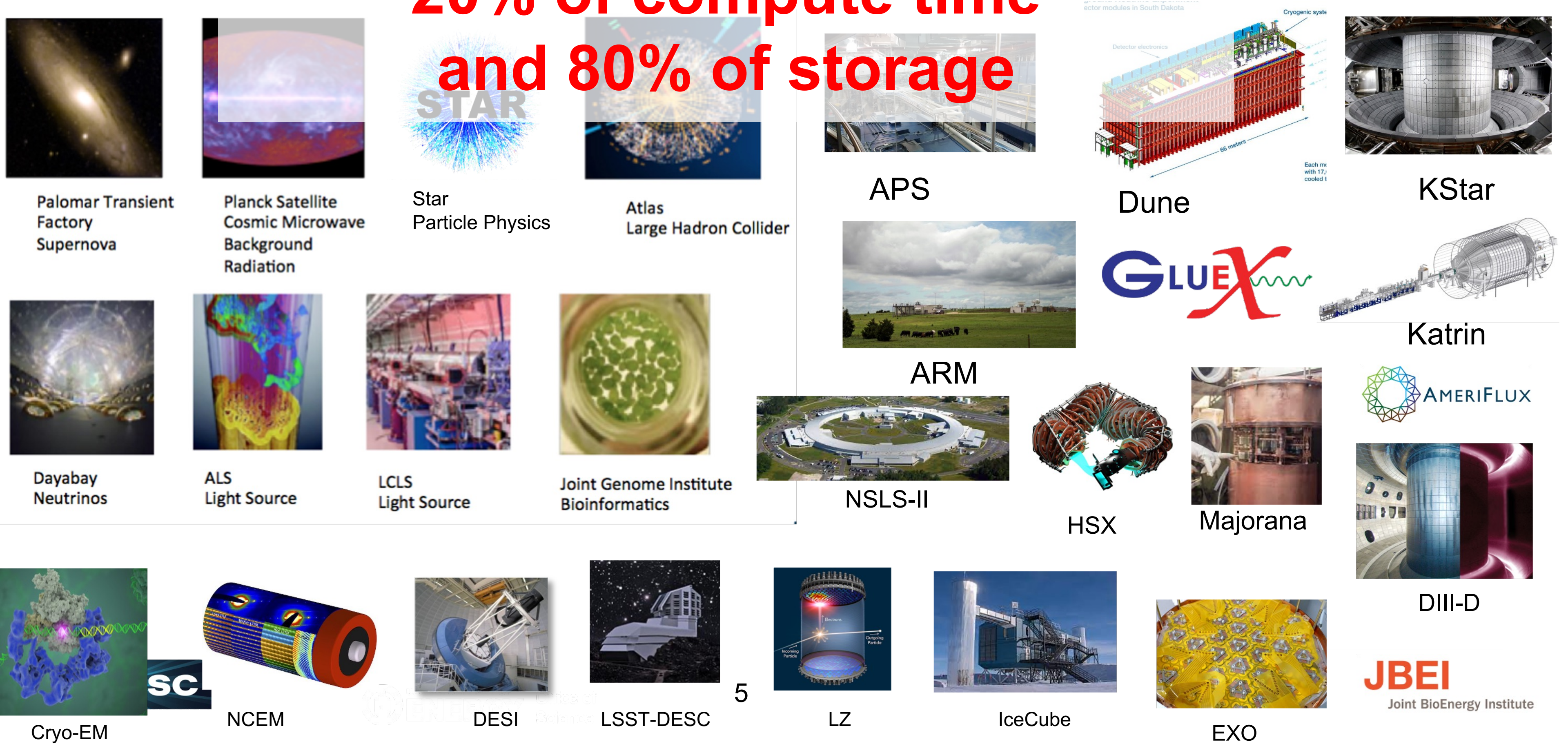

IceCube


EXO

JBEI
Joint BioEnergy Institute

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

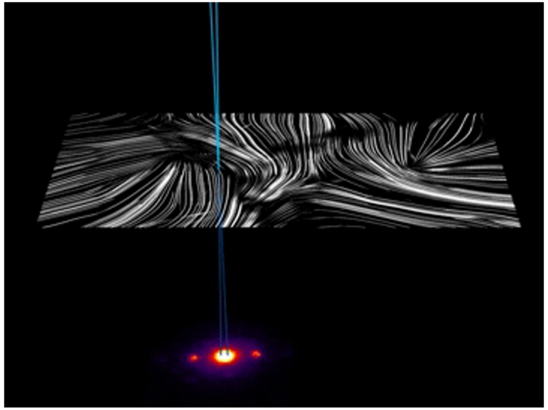
roughly 30% of NERSC users, 20% of compute time and 80% of storage



New experiment technology creates new data challenges

National Center for Electron Microscopy (NCEM) at Berkeley Lab

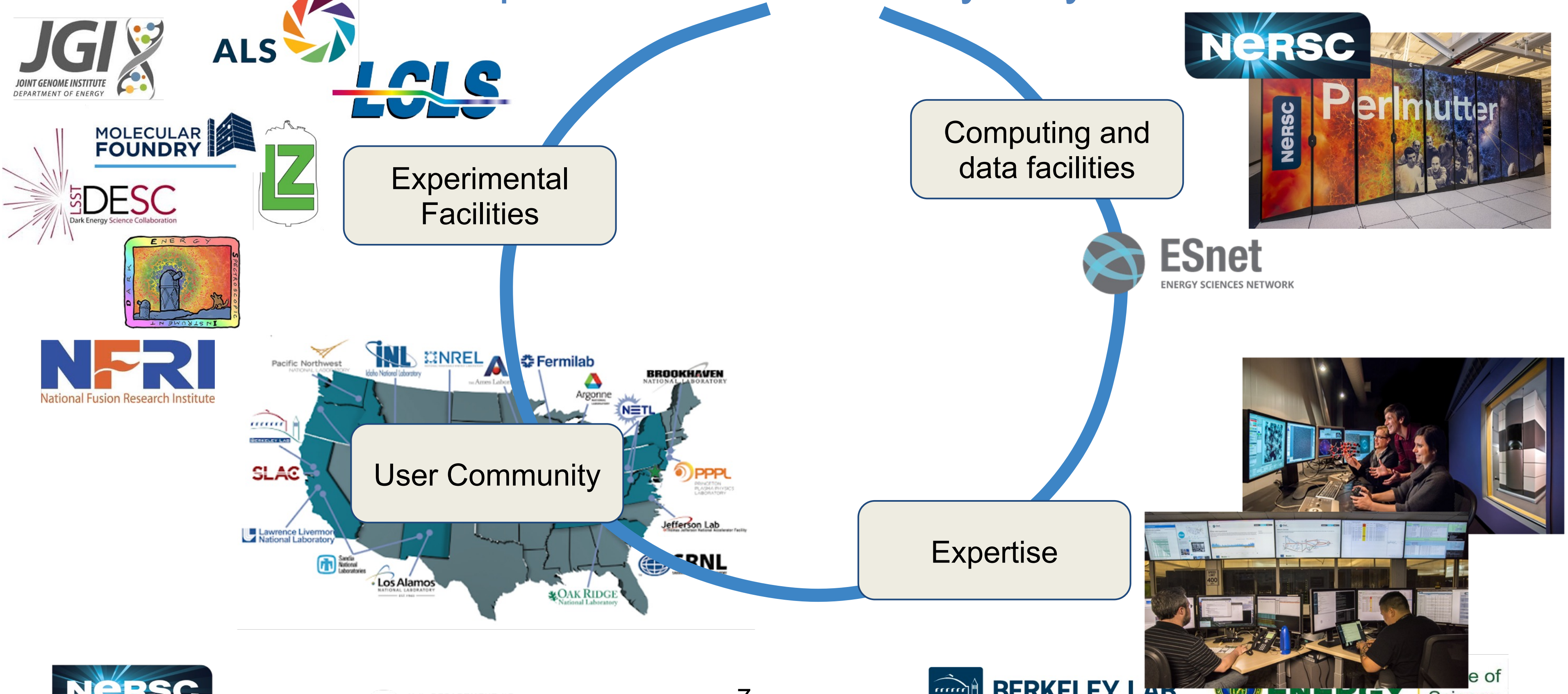
- How does the structure of batteries impact their performance? Can nanocrystals be used to store carbon dioxide?



NCEM is developing new detectors for 4D scanning transmission electron microscope

- 1kx1k pixel scan captures 700 GB in 15 seconds
- Needs HPC-scale computing to analyze data while user is at the microscope

The Superfacility concept: connecting experiment and compute facilities with the expertise and community they need for success

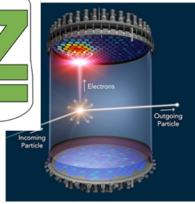


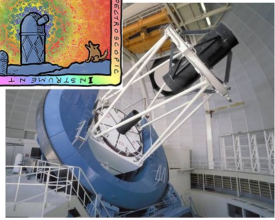
The Superfacility ‘project’ coordinated our work to support the Superfacility Model

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**





AMCR
SciData

Three principles behind our project approach

Integrated requirements from multiple teams

Integrated work across many groups at Berkeley Lab

Scalable to full user base

Scalable to supercomputer capabilities

Sustainable software design model

Sustainable user support model

Superfacility work is driven by science needs: close partnership with science engagements

User-facing tools and policies

- Scalable analysis code, via NESAP
- Outreach and documentation
- Policies
- Jupyter

Scheduling and Middleware

- Reservations and real-time scheduling
- API into NERSC for automation
- Federated Identity
- Spin
- Workflow Resiliency

Science requirements

Automation and Networking

- Software-Defined Networking (SDN)
- Self-managed Systems
- SENSE, for API-based WAN network provisioning and control

Data Management

- External and Internal Data Movement
- Data and PI Dashboards
- HDF5

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.

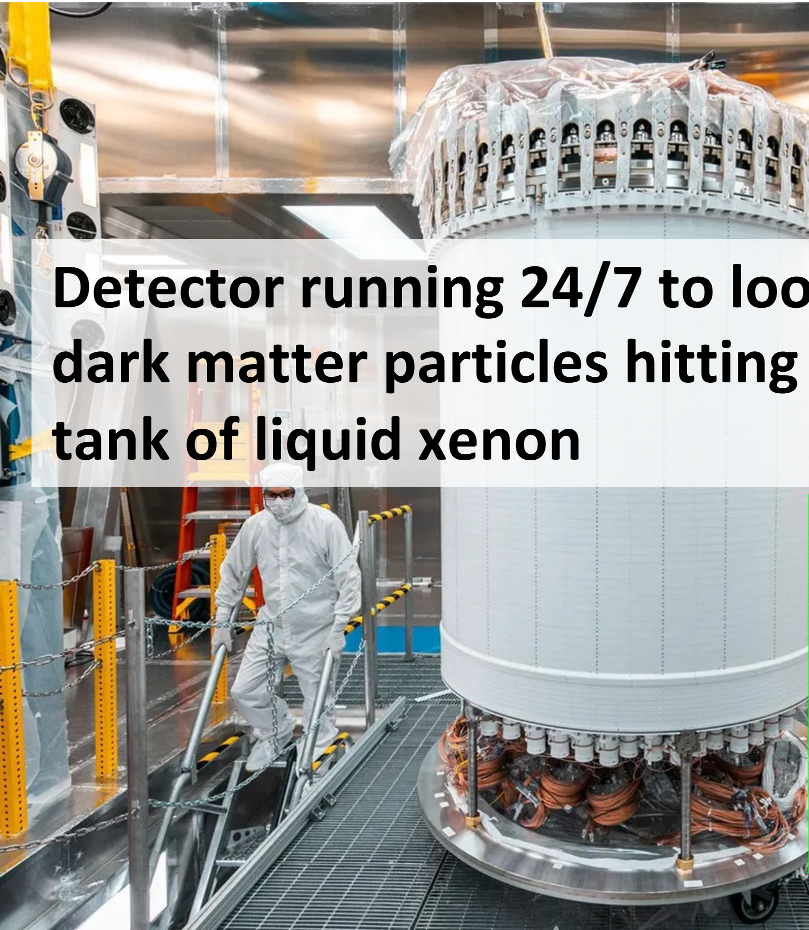
- *Access HPC file systems and networks*
- *Use public or custom software images*
- *Orchestrate complex workflows*
- *Secure, scalable, and managed*

docker RANCHER kubernetes

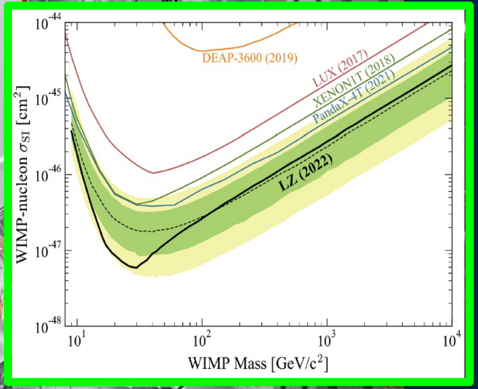
Some projects using Spin:

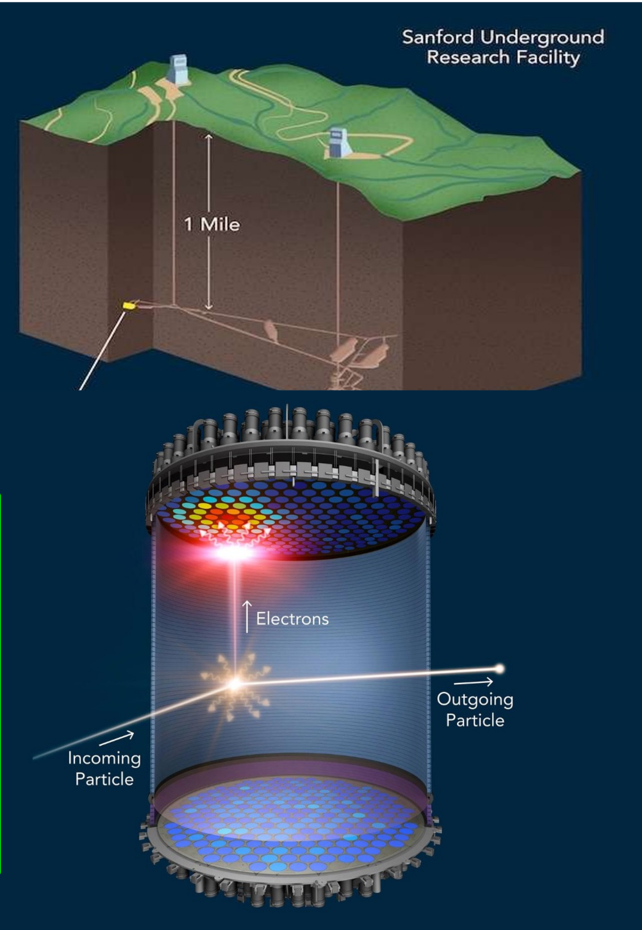
DARK ENERGY SPECTROSCOPIC INSTRUMENT	Track and compare analyses of nightly sky surveys	science gateway
ESS-DIVE	Classify and store reusable earth sciences data	data repository
JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY	Manage production genomic workflows and data at scale	science gateway
LZ	Process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

LZ uses NERSC to search for dark matter particles

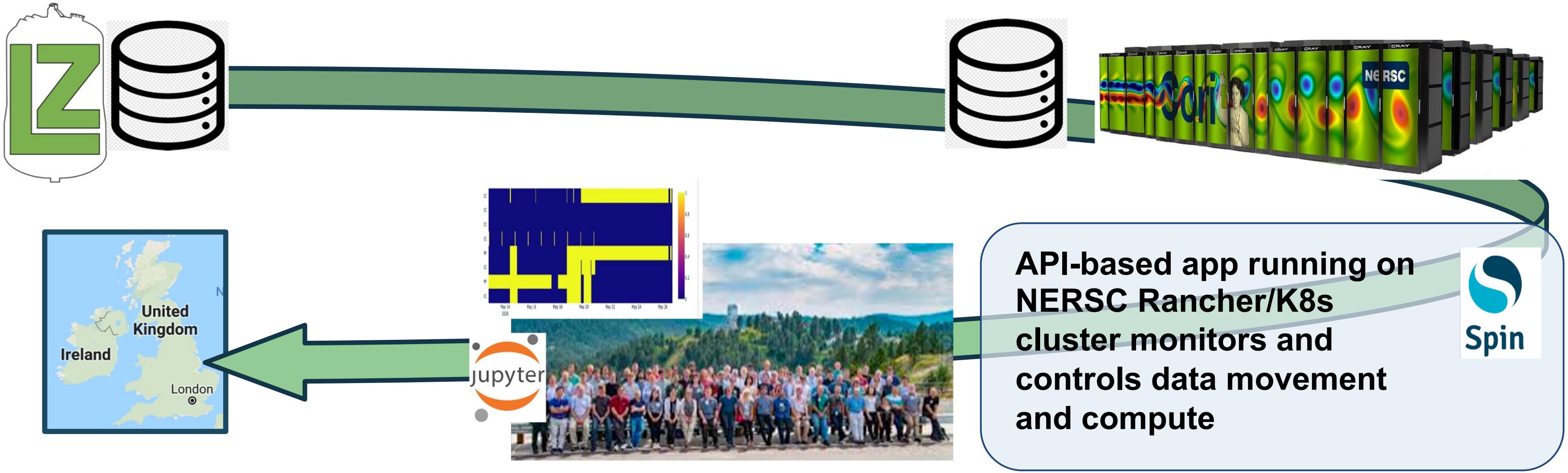


Detector running 24/7 to look for dark matter particles hitting a tank of liquid xenon





LZ uses NERSC to search for dark matter particles



Key needs: Automated, continuous analysis and data movement between data centers, plus offline simulation and analysis by large collaboration

- API + cloud-inspired services + real-time computing = *smooth movement of data and monitoring of detector health*

Q. Riffard (LBNL), ME Monzani (SLAC) and the LZ team

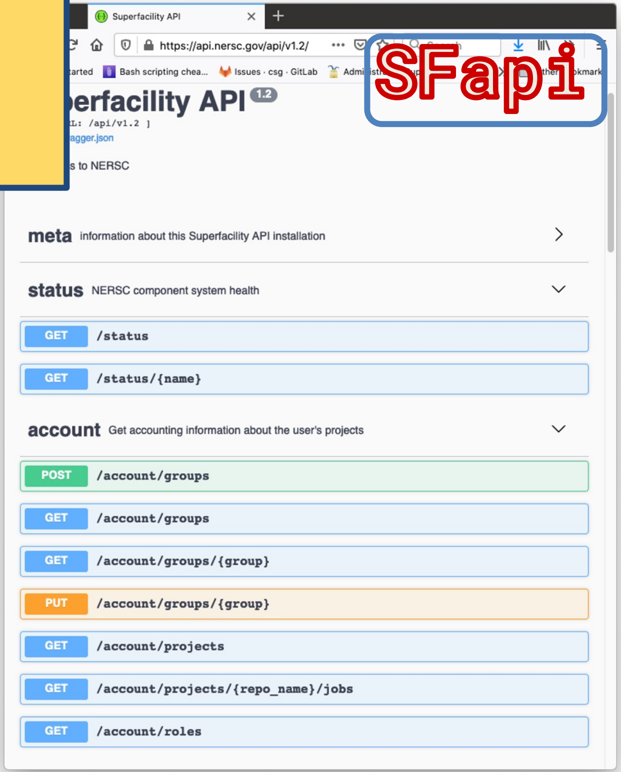
Machine-readable supercomputers: the Superfacility API

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

- A unified programmatic approach to accessing NERSC
- REST API with json input/output
- Standards-based authentication
- Aligned with with CSCS (Swiss National Computing Center) API as much as possible.
- End user docs and examples: https://docs.nersc.gov/services/sfapi/

Since release in 2022

- 27 non-staff users made clients
- members of 40 different non-staff projects.



https://api.nersc.gov/V1.2

Credit: Bjoern Enders (NERSC)

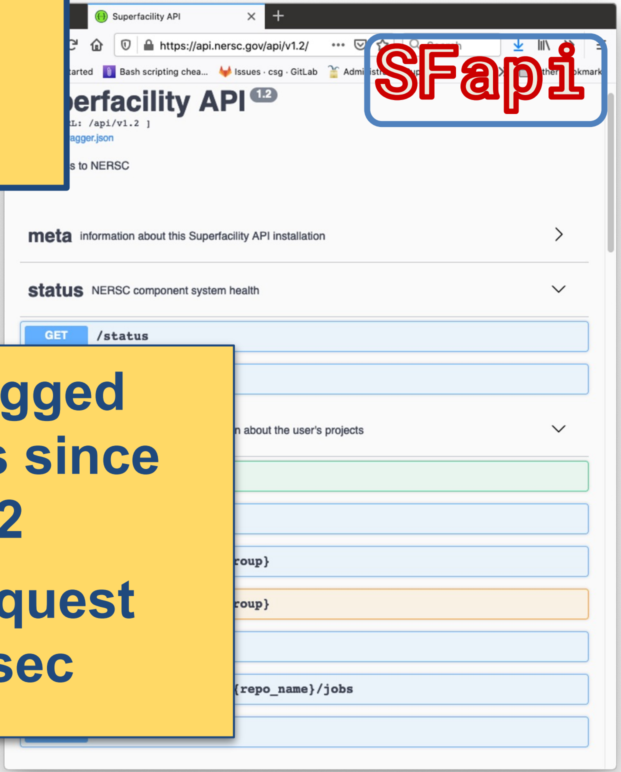
Machine-readable supercomputers: the Superfacility API

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

- A unified programmatic approach to accessing NERSC
- REST API with json input/output
- Standards-based authentication
- Aligned with with CSCS (Swiss National Comput
  API as much as possible.
- End user docs and examples: https://docs.nersc.gov/services/sfapi/

Since release in 2022

- 27 non-staff users made clients
- members of 40 different non-staff projects.



~ 12M logged requests since May 2022

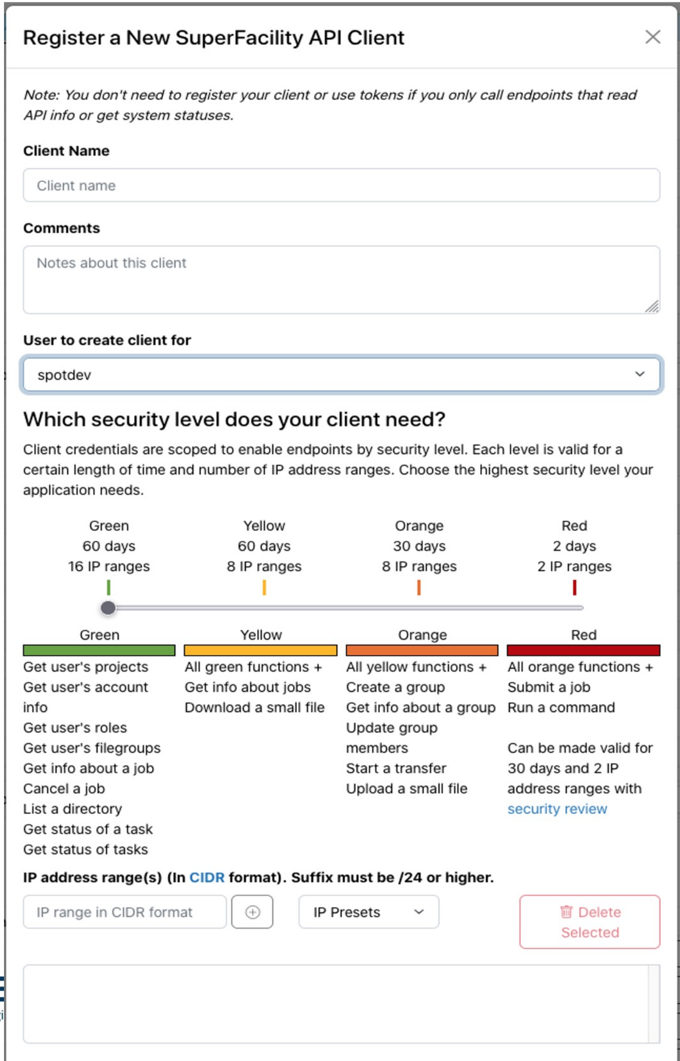
= one request every 2 sec

https://api.nersc.gov/V1.2

Credit: Bjoern Enders (NERSC)

Opening up NERSC to API calls took careful consideration

- Conducted multiple UX reviews
  - An analysis from the user point of view → made changes for functionality and ease of use
- Conducted multiple security reviews
  - Included both API architecture and new OpenID-based authentication
  - Authentication model requires strict credential lifetimes - need to enforce MFA
  - Each endpoint+method is assessed individually on its threats. The assessment determines the max number of IPs and maximum lifetime of a client that has this endpoint in its scope.

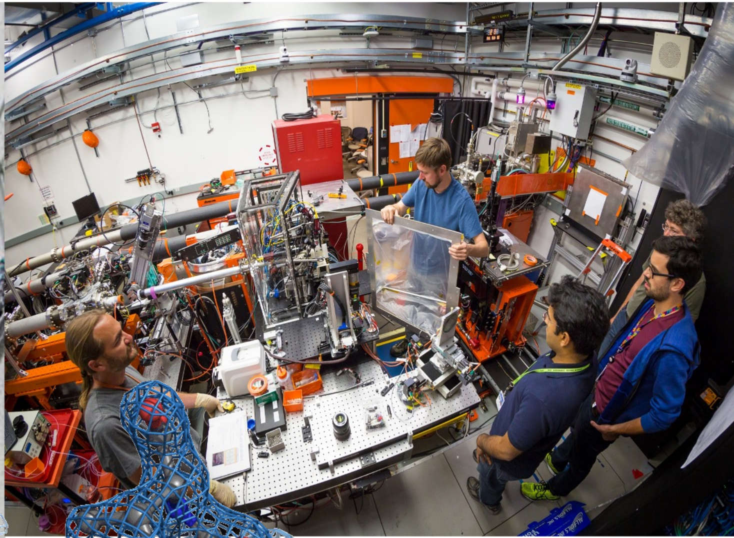


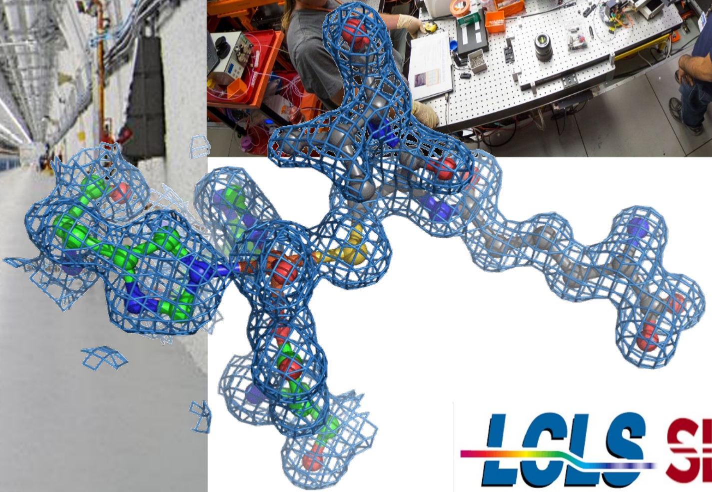
Credit: Bjoern Enders (NERSC)

LCLS is using NERSC for realtime collaborative distributed data analysis

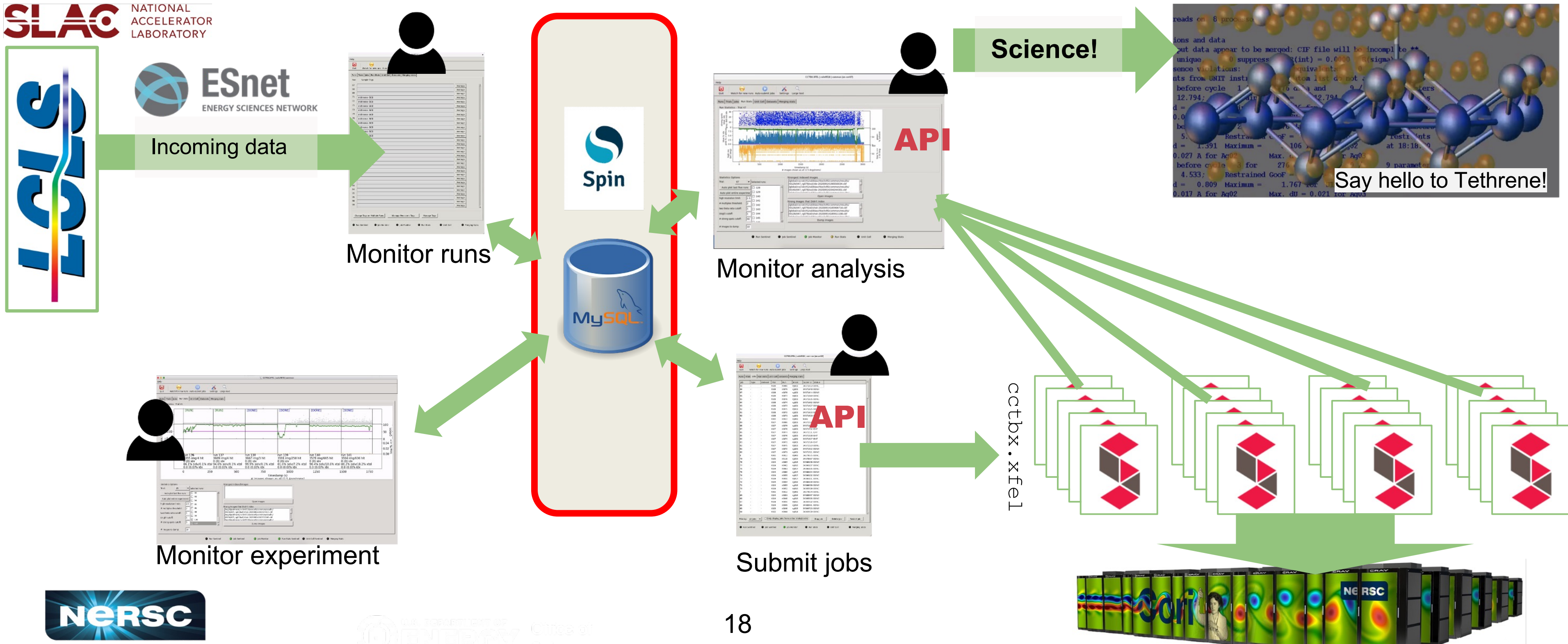
Linac Coherent Light Source produces up to 10PB of data per experiment to create "molecular movies"

- **How does photosynthesis happen?**
- **How do drugs dock with proteins in our cells?**
- **Why do jet engines fail?**



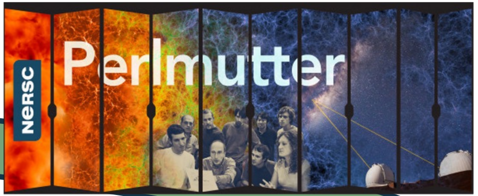


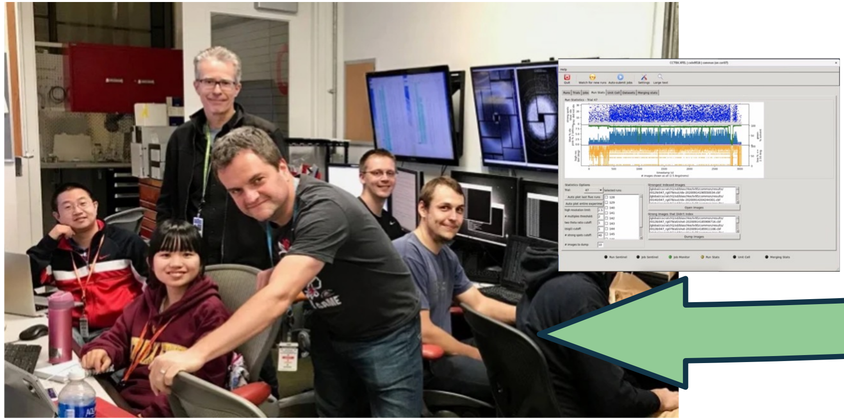
LCLS is using NERSC for collaborative distributed Data Analysis with Spin and the SFapi



Credit: J. Blaschke, F. Wittwer (NERSC) and the ExaFEL & LCLS team

LCLS is using NERSC for realtime collaborative distributed data analysis





API-based app running on NERSC Rancher/K8s cluster monitors and controls data movement and compute

Spin

API

Key needs: Automated, fast turnaround, large-scale data analysis

- API + cloud-inspired services + real-time computing = *results within minutes of data taking*

Credit: J. Blaschke, F. Wittwer (NERSC) and the ExaFEL & LCLS team

Jupyter: supercharge interactive supercomputing

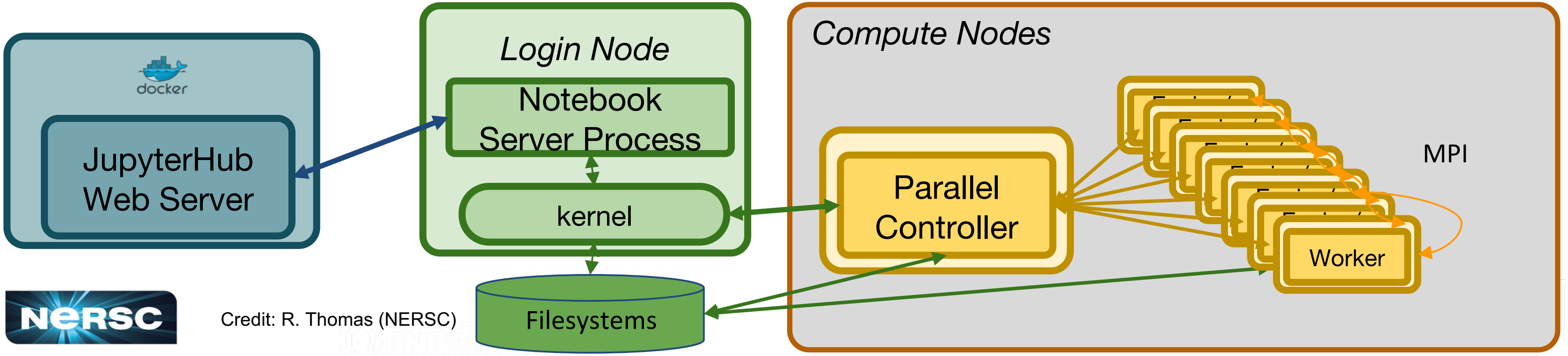
We have deployed an HPC-aware Jupyter service:

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

User quote: "*The 3 most important things in life: food, shelter and Jupyter… everything else is optional.*"

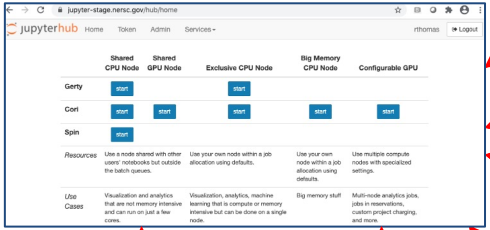
Interactive supercomputing: Jupyter Notebook + HPC Workers

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



Credit: R. Thomas (NERSC)

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

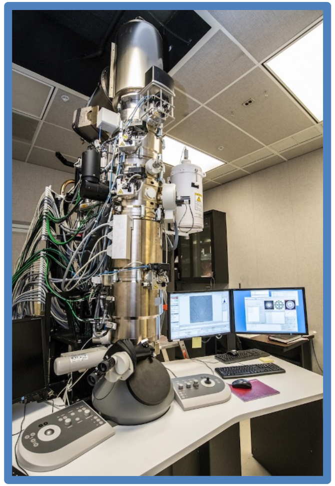
Do you have access to a reservation?
Is the reservation active now?

SFapi

Microservices
Service-oriented architecture

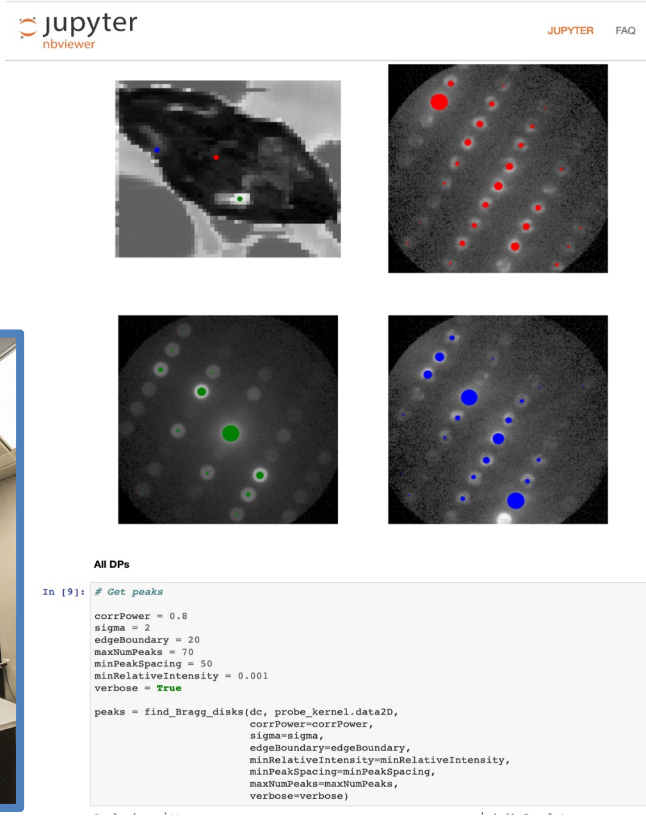
NCEM is using Jupyter and Dask for interactive exploration and analysis of EM images

- Dask is a powerful backend to manage remote workers on a cluster via Python notebooks.
- LBL team 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

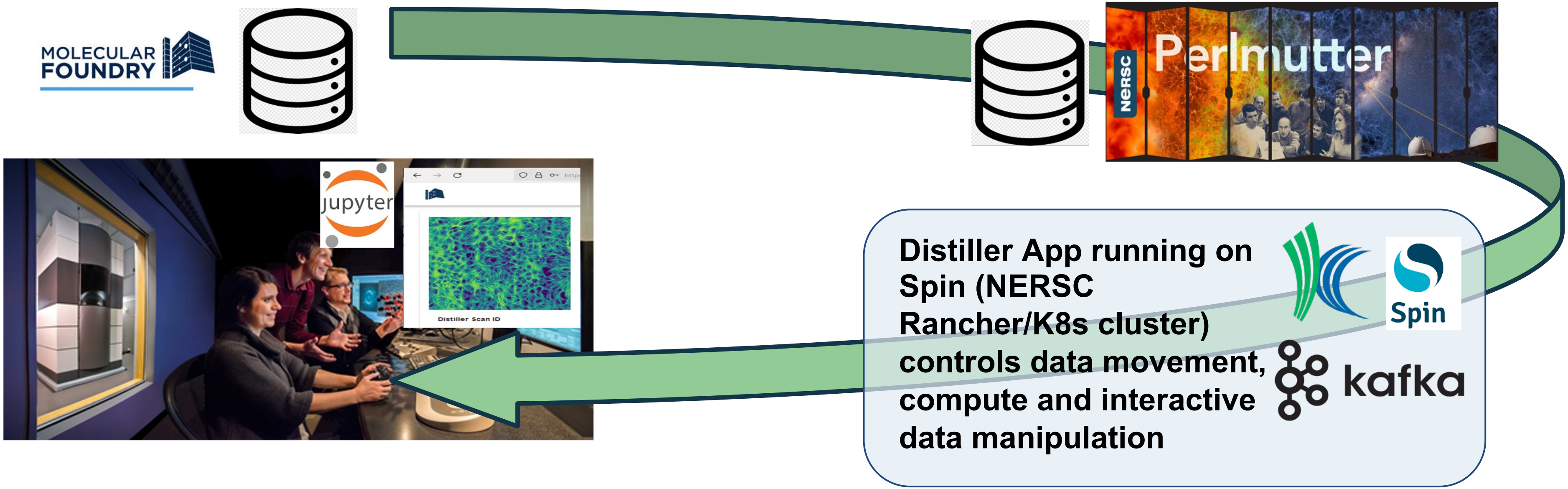


- NCEM: Serial processing of 4D image arrays in numpy - Parallelize it!
- Achieved **20-50x speedup** on NCEM Py4DSTEM Notebooks

Credit: S. Cholia (LBL) and the NCEM team



NCEM is using NERSC for realtime data analysis



Key needs: Automated, fast turnaround, large-scale data analysis

- API + cloud-inspired services + real-time computing + dynamically configured network + Jupyter-based analysis on HPC = *results within minutes of scan*

Resilience is a challenge for experiment sciences

Systems cannot guarantee 24/7 uptime

- Security patches, facility power work, components/power failing…
- IO impacts from ”bad” workload, network contention…

Commercial cloud providers have the same outages, but they are hidden from users by spare capacity and application design.

Biggest remaining challenge: Robustness / Resilience, especially “soft” outages, e.g. transient I/O or slurm failures

NERSC has worked hard to improve our resilience, and we want to help science teams develop more resilient workflows

- We are now able to keep most of our infrastructure up during power work or routine maintenances
- Rolling updates to deploy software/firmware patches across compute and storage

A truly resilient workflow needs to span multiple computing centers

Attempting to port an established, operational pipeline to another site is very, very hard

Experimental science data analysis pipelines need 24/7/365 HPC resources, which can only be achieved by computing at multiple locations.

We attempted to port workflows from NERSC to a LBNL cluster and discovered all kinds of unexpected pain points

- Access and authorisation control across sites
- Database robustness and mirroring
- Harmonising software libraries across sites
- Container build wrappers for multiple architectures, MPI libs etc
- Orchestration and web service access outside of institution network
- Different filesystem types, interfaces, sizes, permission structures

→ ***If a workflow is not designed with portability in mind, it will be very difficult to use an IRI***

Shared Burden

- Work together between systems and their users
- Containers are great
  - Users should have best practices to make them portable
- Developers should think ahead about cross-facility workflows
  - Rewriting code for each site is hard
  - Infrastructure as code
- Facilities adopt APIs to interact with parts of their system
  - A standard set of calls
    - Checking status & Starting jobs & Transferring data
  - Help eliminate duplication of code
  - Minimal effort to add a new site which uses the standard

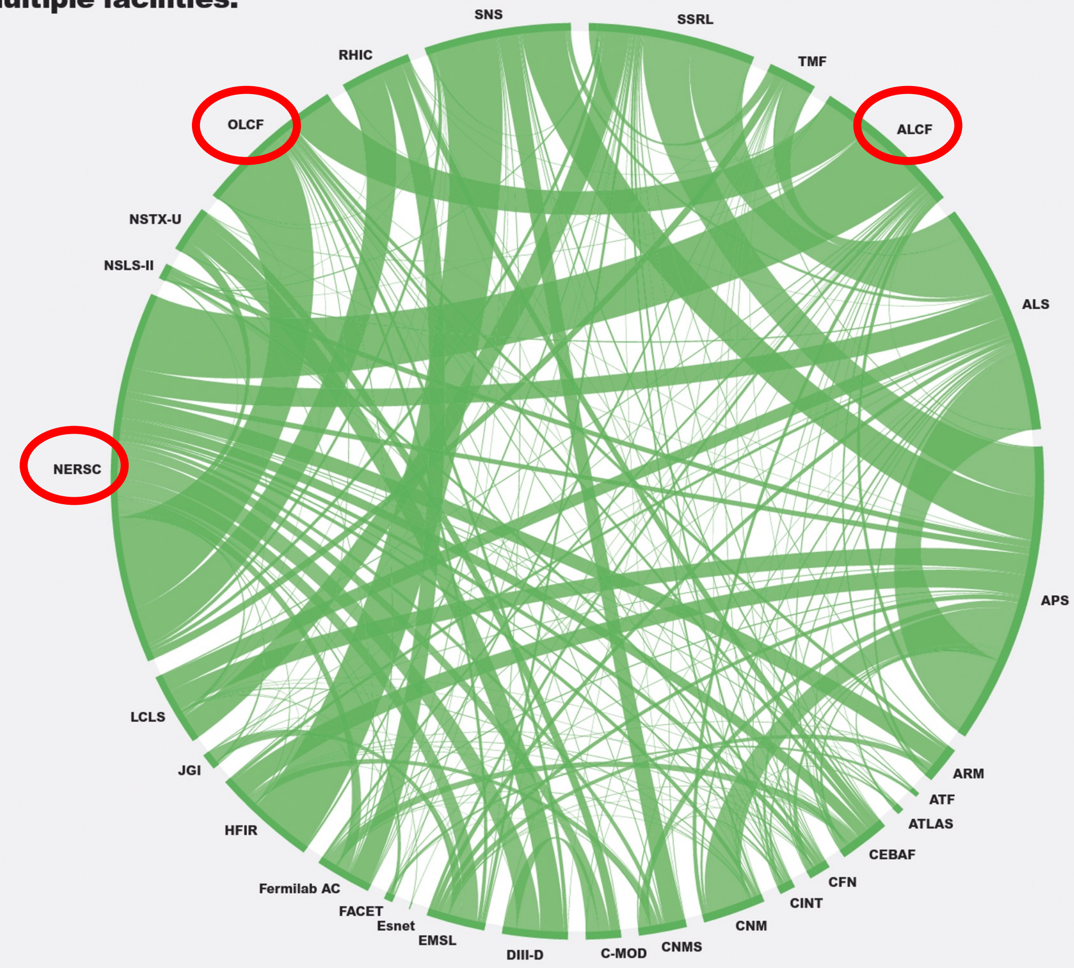
From Nick Tyler and Rob Knop, WORKS21, https://ieeexplore.ieee.org/document/10023937

Each facility provides a unique scientific toolset. Users enhance their research by leveraging capabilities at multiple facilities.

Users of one facility are often users of multiple facilities.

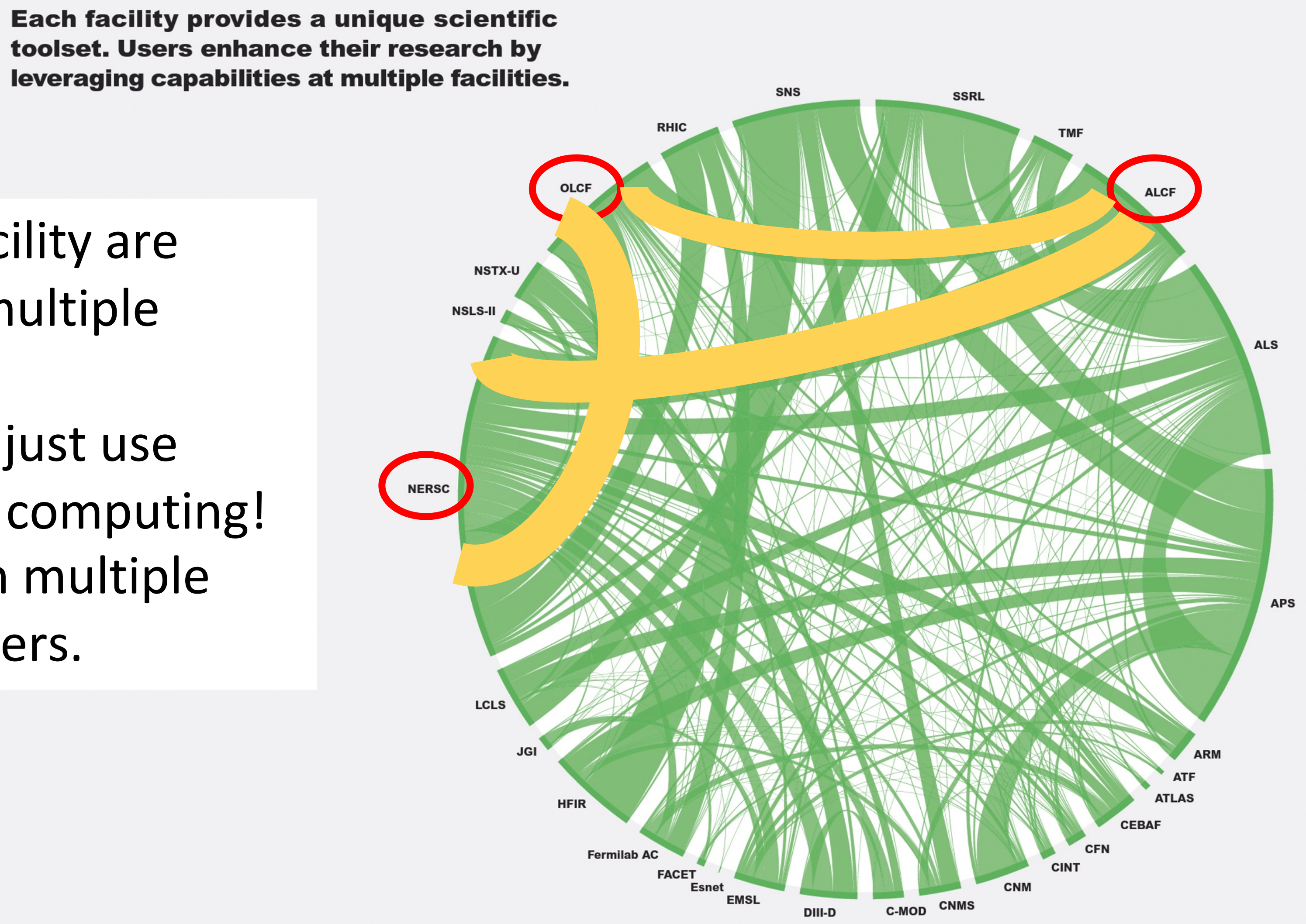
Scientists don't just use NERSC for their computing!

Workflows span multiple computing centers.

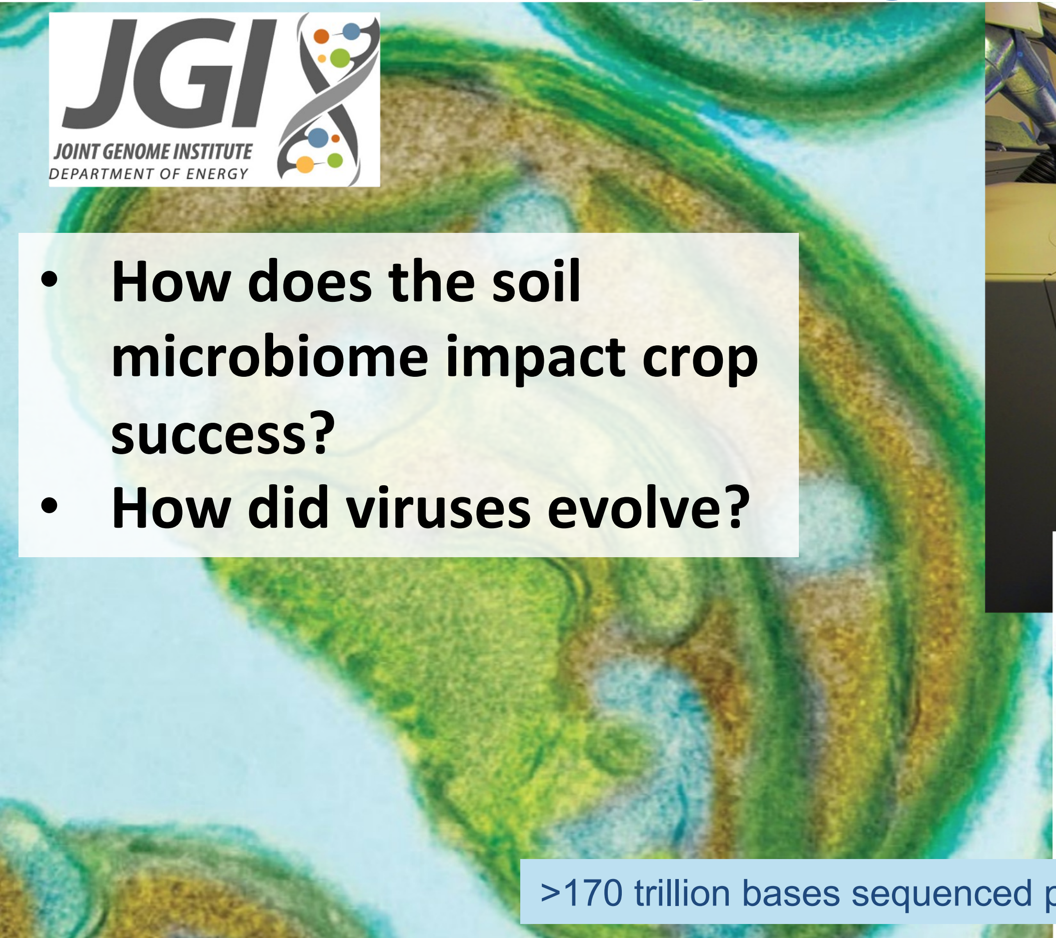


Users of one facility are often users of multiple facilities.

Scientists don't just use NERSC for their computing! Workflows span multiple computing centers.

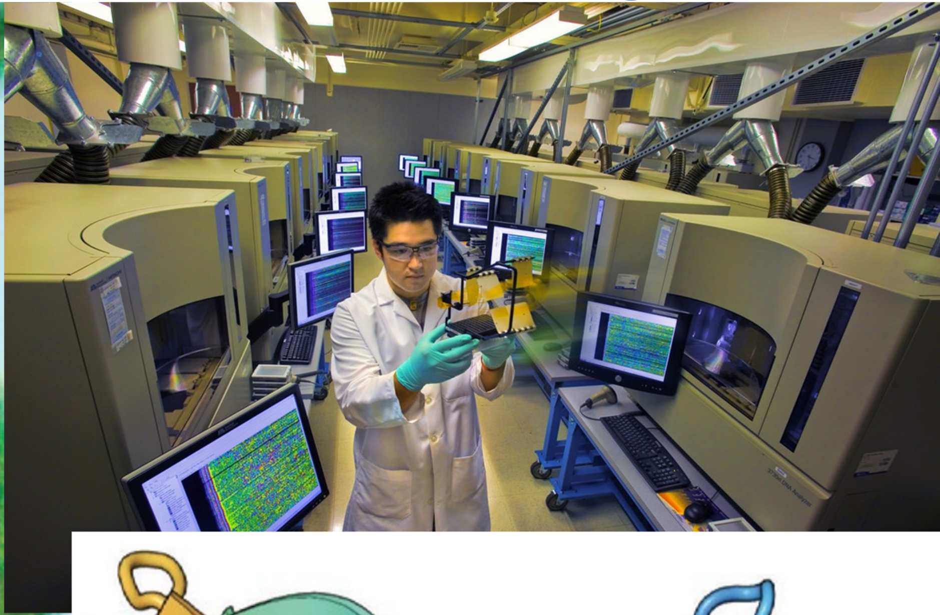


Supercomputing for genome sequencing



JGI
JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

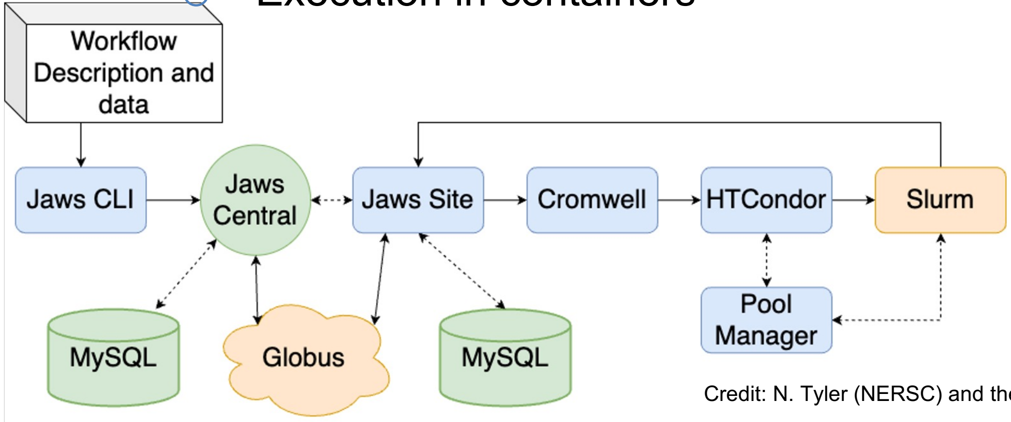
- How does the soil microbiome impact crop success?
- How did viruses evolve?

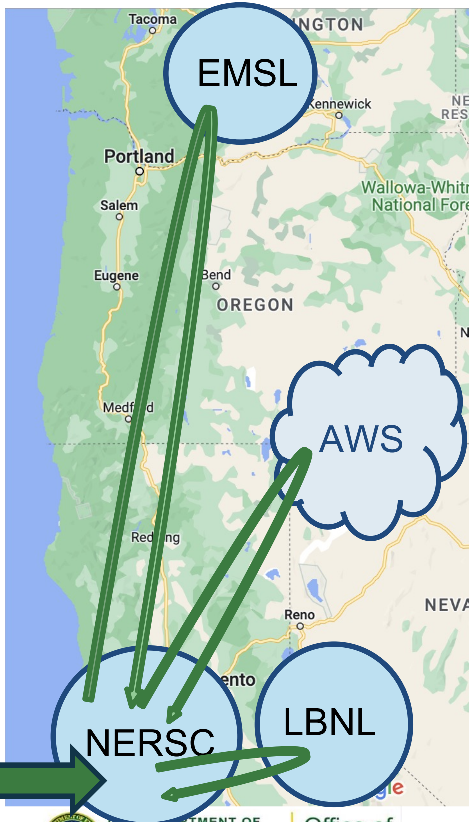


>170 trillion bases sequenced per year, >7PB of archived data, >100,000 users

The JGI has developed cross-site automated workflows

- JGI staff submit workflows defined via WDL (Workflow Description Language), specifies location where analysis should run
- JAWS is WaaS, handles:
  - Data movement to/from a site via Globus
  - Resource allocations via HTCondor and Slurm
  - Execution in containers

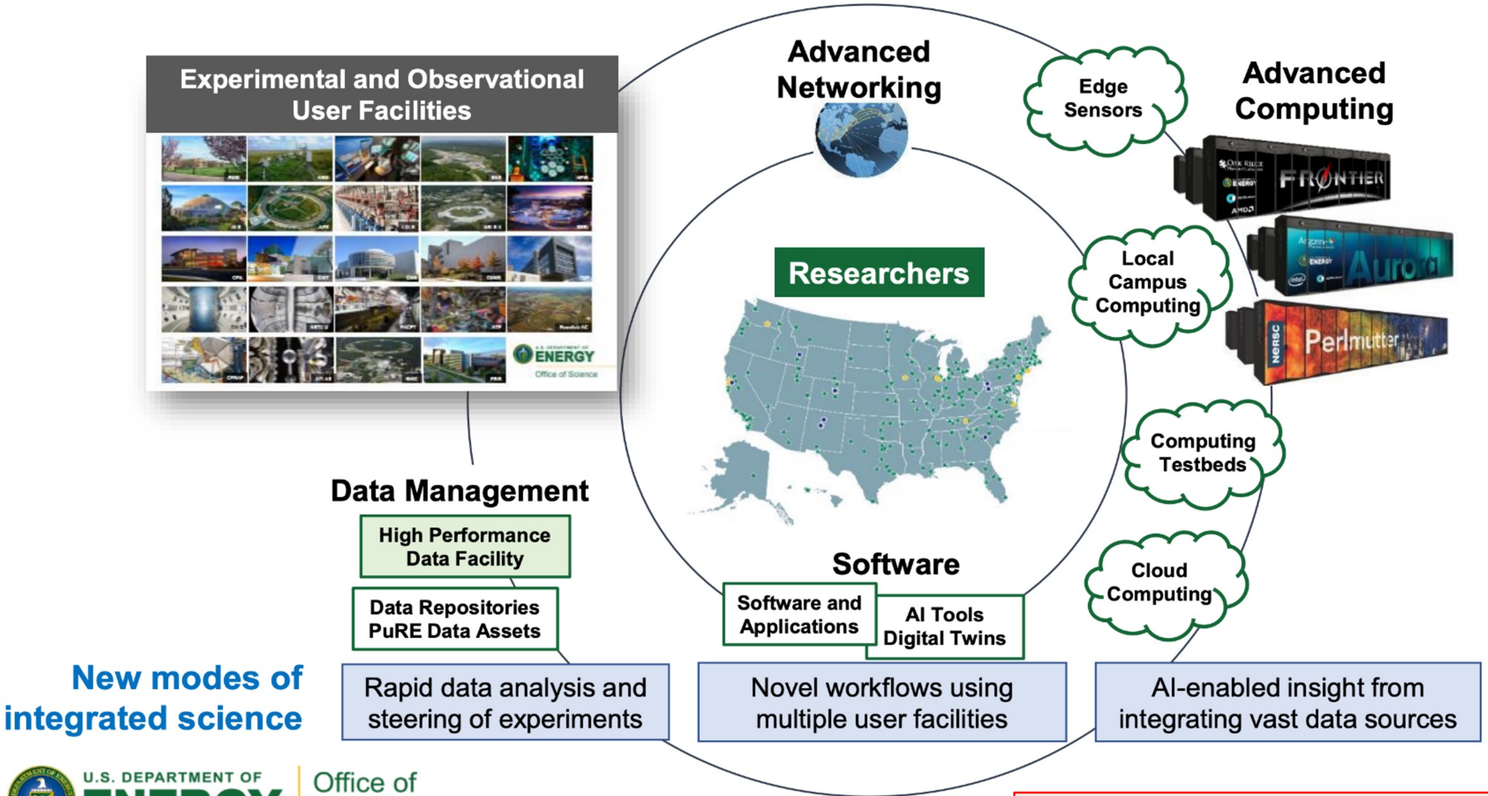




Credit: N. Tyler (NERSC) and the JAWS team at JGI

DOE’s Integrated Research Infrastructure (IRI) Vision:

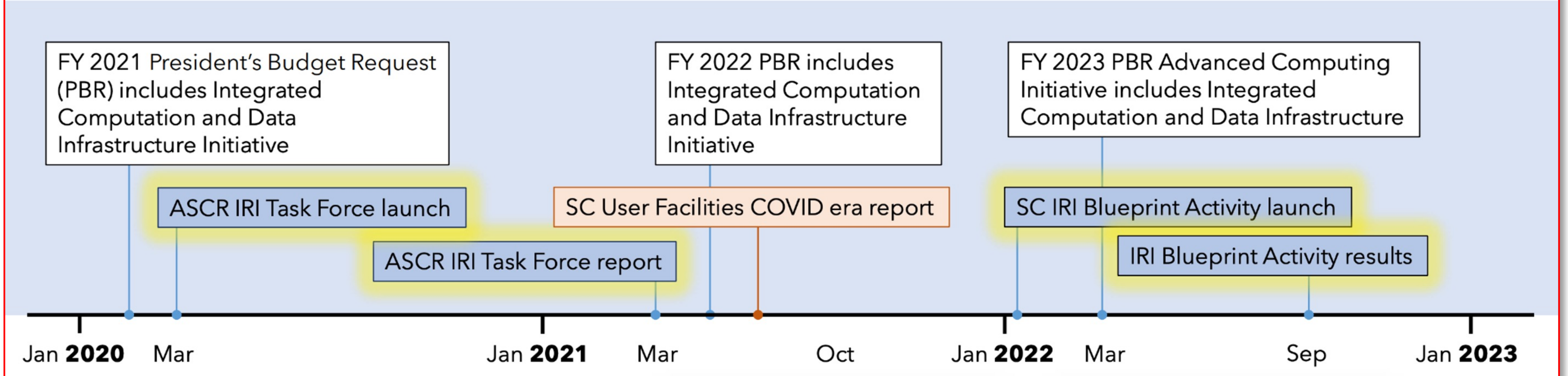
To empower researchers to meld DOE’s world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation



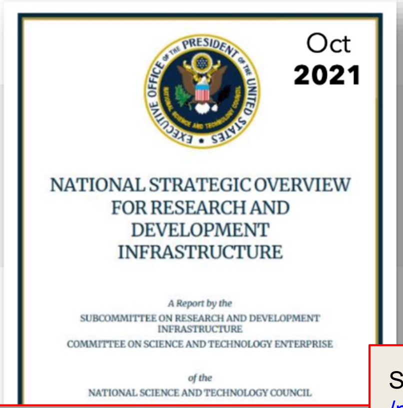
U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Ben Brown, ASCR (https://science.osti.gov/-/media/ber/berac/pdf/202304/Brown_IRI_for_BERAC_202304.pdf)

Timeline of key IRI activities, 2020-22



Integration of instrumentation, data, and computing infrastructure are essential requirements for national R&D objectives



NATIONAL STRATEGIC OVERVIEW FOR RESEARCH AND DEVELOPMENT INFRASTRUCTURE

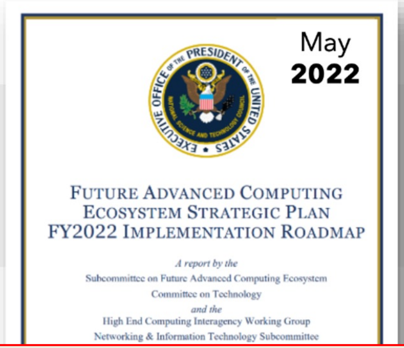
A Report by the

SUBCOMMITTEE ON RESEARCH AND DEVELOPMENT INFRASTRUCTURE

COMMITTEE ON SCIENCE AND TECHNOLOGY ENTERPRISE

of the

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL



FUTURE ADVANCED COMPUTING ECOSYSTEM STRATEGIC PLAN FY2022 IMPLEMENTATION ROADMAP

A report by the

Subcommittee on Future Advanced Computing Ecosystem

Committee on Technology

and the

High End Computing Interagency Working Group

Networking & Information Technology Subcommittee

U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Ben Brown, ASCR (https://science.osti.gov/-/media/ber/berac/pdf/202304/Brown_IRI_for_BERAC_202304.pdf)

IRI Blueprint Activity Key Results

We now possess a reference framework to inform a coordinated, SC-wide strategy for IRI.

The key organizing elements of the IRI Framework are Science Patterns and Practice Areas:

> **IRI Science Patterns** that represent integrated science use cases across DOE science domains and

> **IRI Practice Areas** that will support the realization of a DOE-integrated IRI ecosystem.

U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Ben Brown, ASCR (https://science.osti.gov/-/media/ber/berac/pdf/202304/Brown_IRI_for_BERAC_202304.pdf)

ASCR recently put out a call for proposals for a new user facility - the High Performance Data Facility (HPDF)

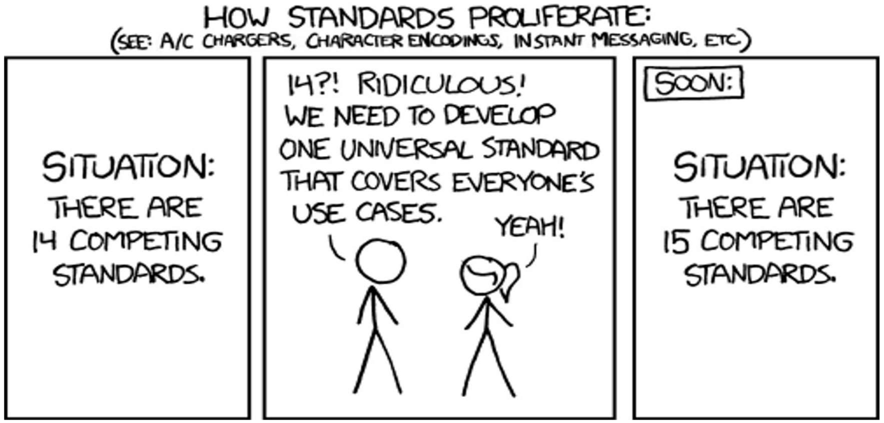
- "a new scientific user facility specializing in advanced infrastructure for data-intensive science."
- "The mission of the HPDF will be to enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools."
- "The facility will be designed to dynamically configure computation, network resources, and storage to access data at rest or in motion, supporting the use of well-curated datasets as well as near real-time analysis on streamed data directly from experiments or instruments."

HPDF will be a key component of a truly integrated research infrastructure.

Making cross-site workflows possible/simple will require a lot of work, investment and cooperation from all computing centers

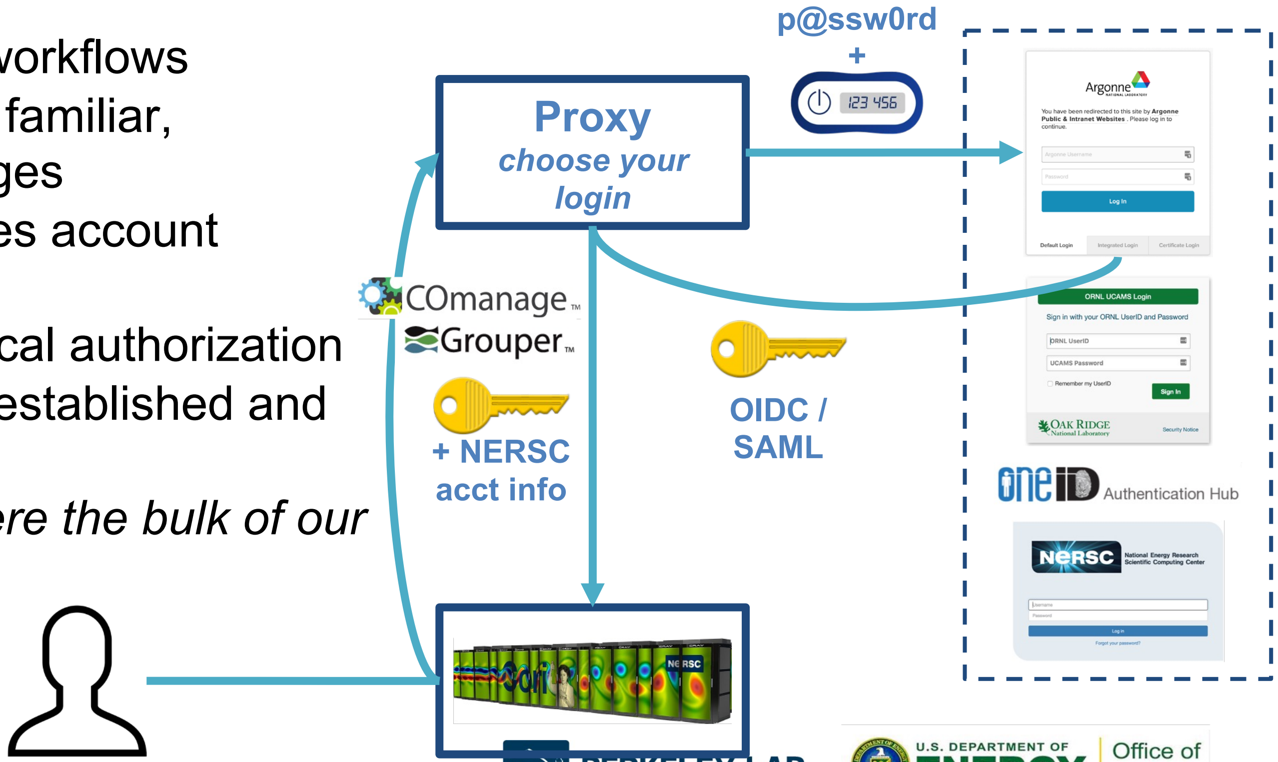
- **Security**: identity and access management across facilities
- **Data management** across multiple sites
- **Scheduling** and resource management, to place compute tasks with the right resources at the right time
- **Portability and scalability** from laptop to supercomputer
- Community agreement on the **interfaces** that will enable portability across sites

Sociological challenges will outweigh the technical



Federated Identity (FedID) at NERSC 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
- 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*



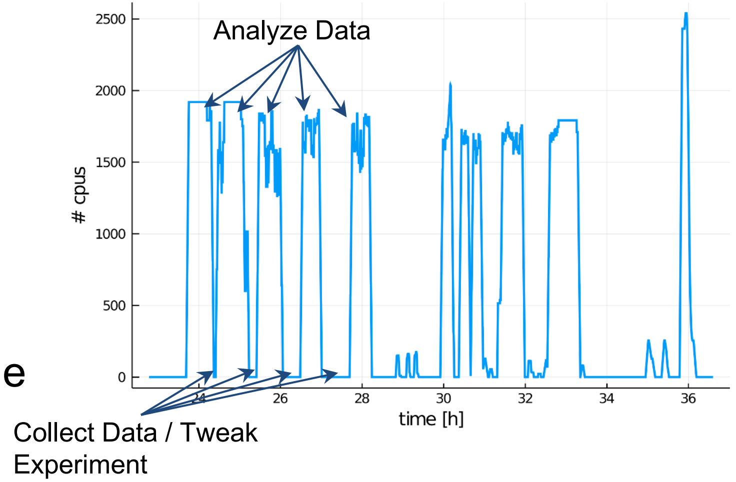
NERSC
BERKELEY LAB Bringing Science Solutions to the World
U.S. DEPARTMENT OF ENERGY Office of Science

Scheduling an urgent workload while maintaining high utilization is challenging

- NERSC typically has thousands of running jobs
- Queue frequently 10x larger (10,000 - 20,000 eligible jobs)
- "Normal" job backlog up to 10 days long

How do we make room for urgent compute requests from experiment teams?

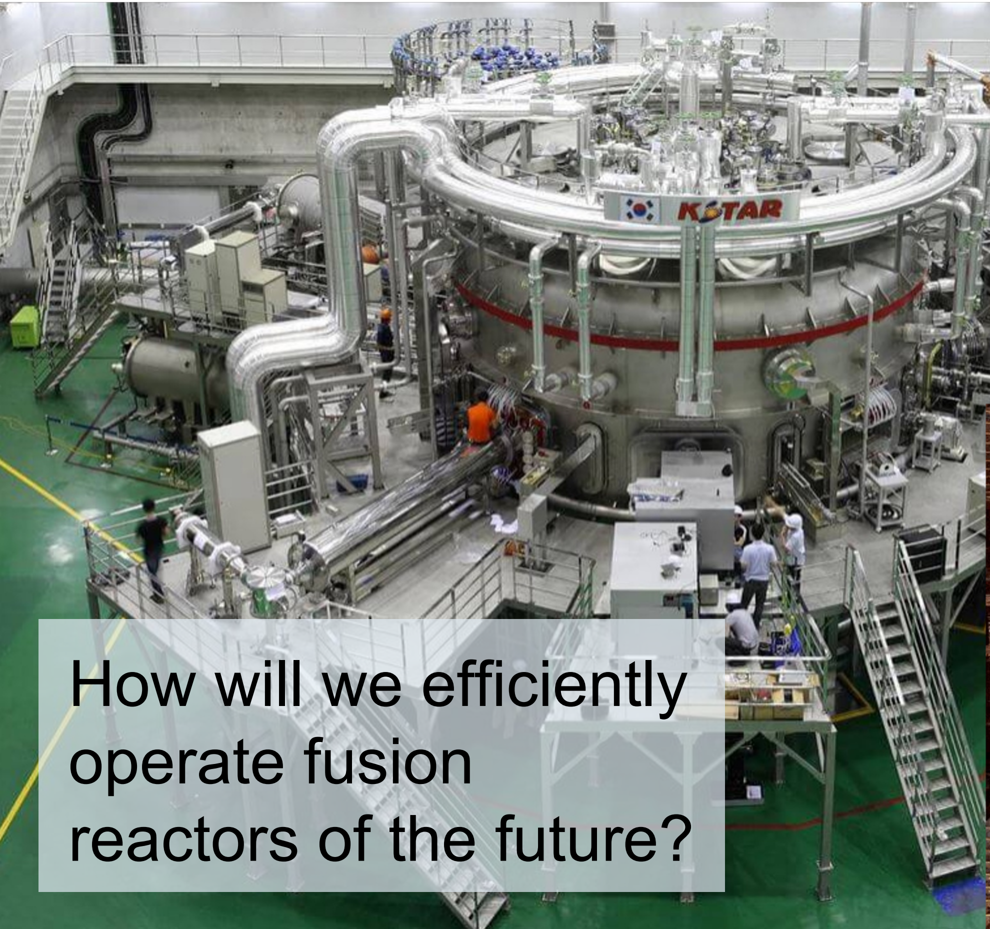
- Realtime queue for small urgent compute
  - Dedicated nodes + high priority
- Reservations for experiment shifts
- Preemptible jobs to fill gaps
  - NERSC funded this capability in Slurm 20.02
  - Investing in checkpointing technology to provide preemptible workload

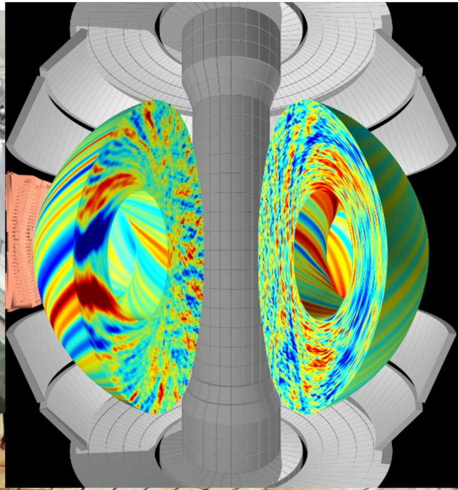


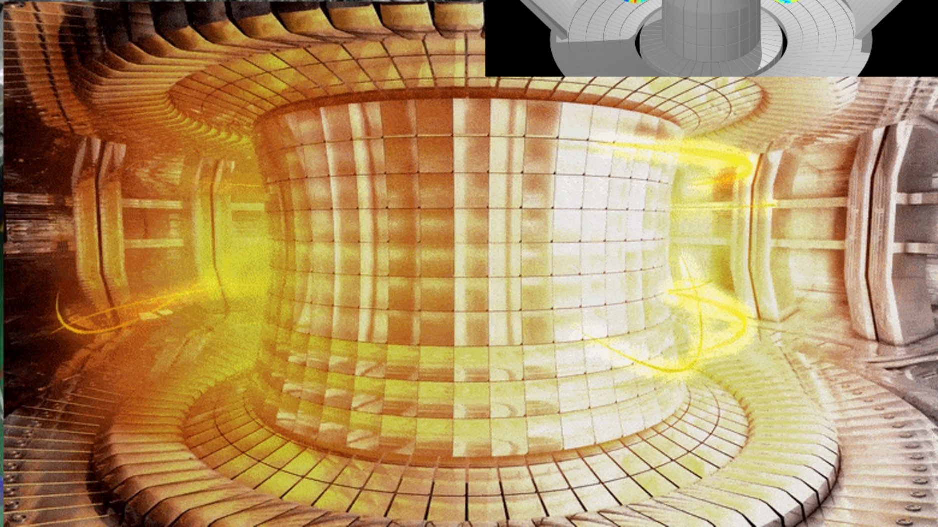
Scheduling work across multiple sites will be a significant challenge

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

Supercomputing for Fusion Energy

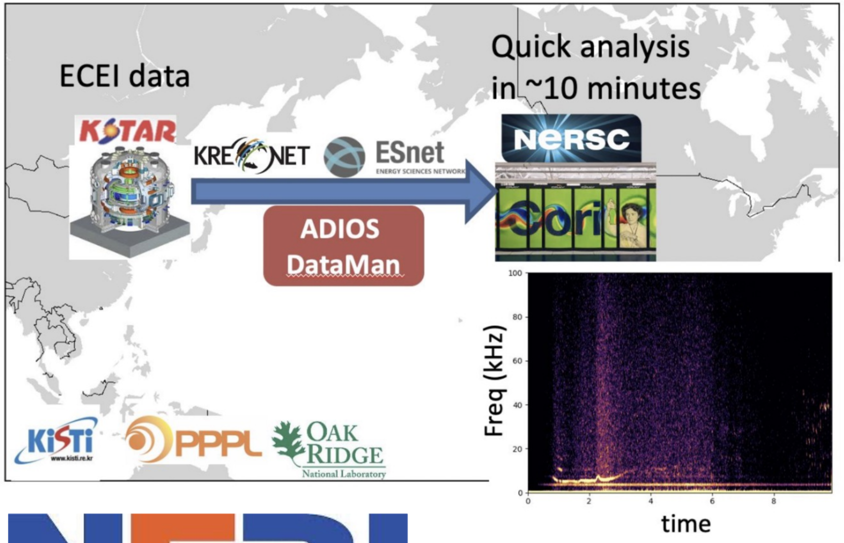






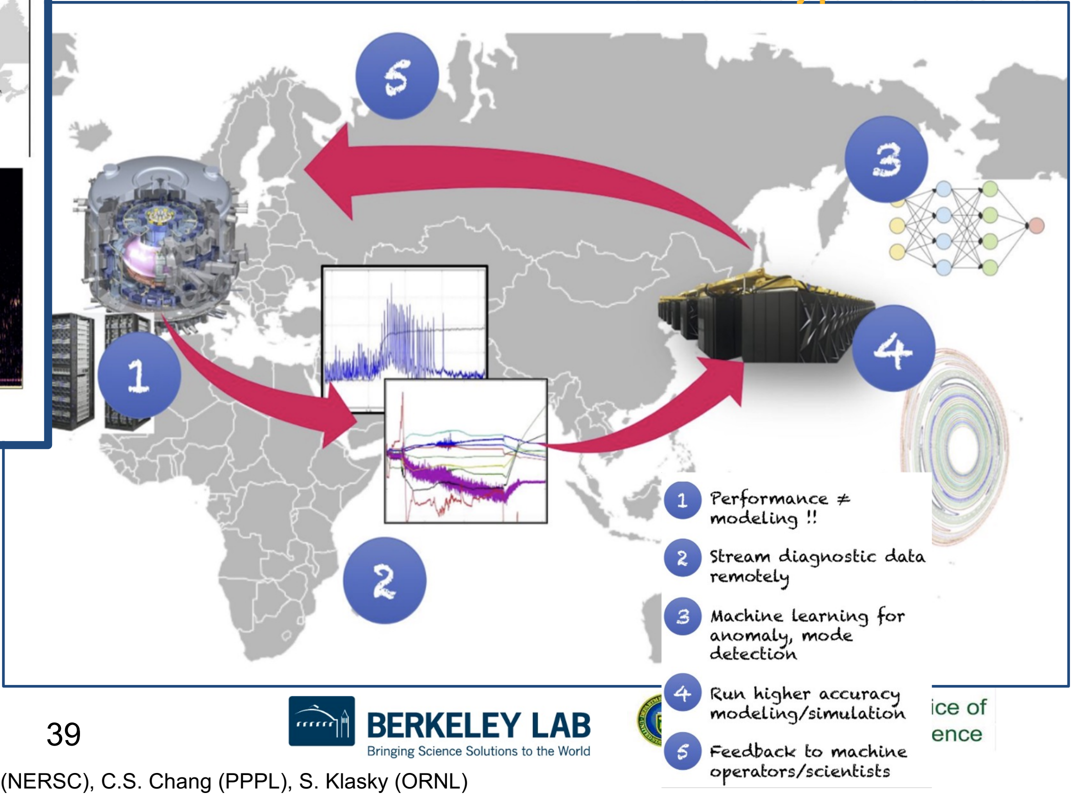
How will we efficiently operate fusion reactors of the future?

An international superfacility



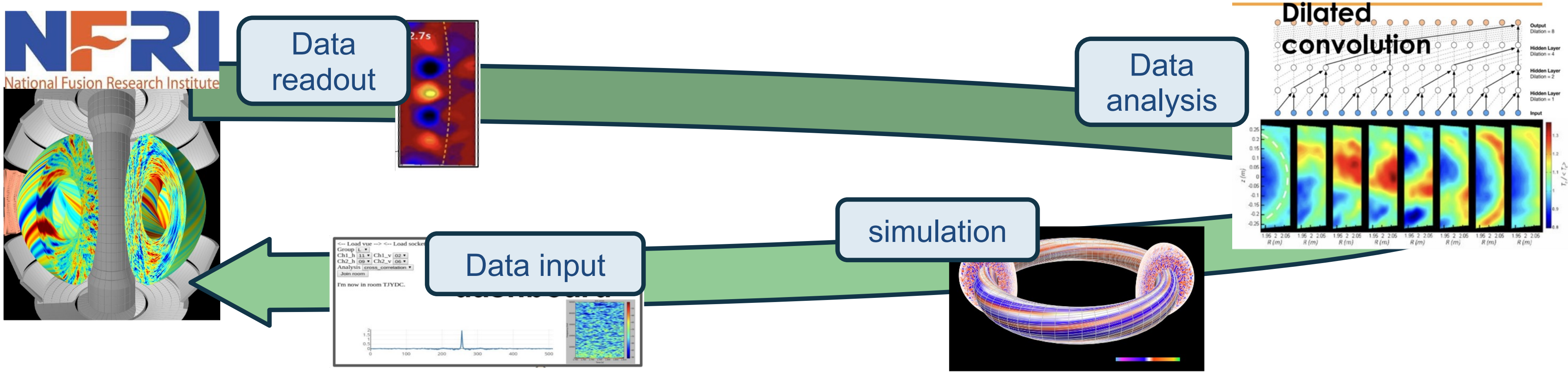
NFRI
National Fusion Research Institute

iter
china eu india japan korea russia usa



R. Kube (PPPL), J. Choi (ORNL), J. Wang (ORNL), L. Stephey (NERSC), C.S. Chang (PPPL), S. Klasky (ORNL)

Realtime simulation + data analysis to control fusion reactors

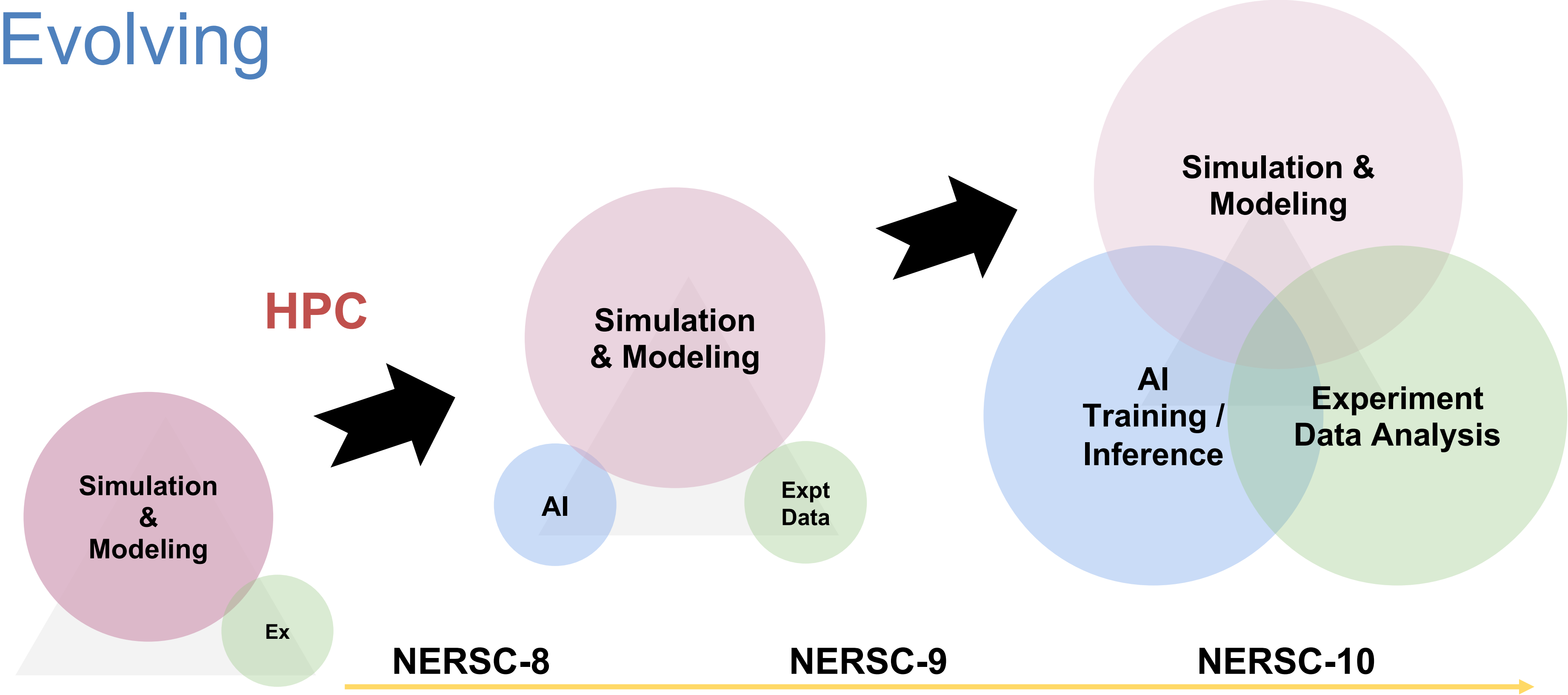


Key needs: Automated, fast turnaround, large-scale data analysis coupled with simulation based on data readout from tokomak

- API + cloud-inspired services + real-time computing + simulation + AI-based data analysis = *scientists can update magnetic field parameters within minutes of a plasma shot*

R. Kube (PPPL), J. Choi (ORNL), J. Wang (ORNL), L. Stephey (NERSC), C.S. Chang (PPPL), S. Klasky (ORNL)

HPC Facility Workload Balance is Evolving

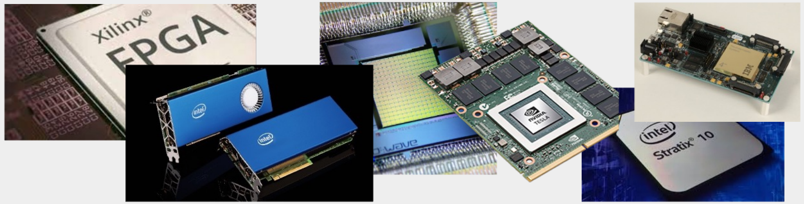


Credit: Hai Ah Nam (NERSC)

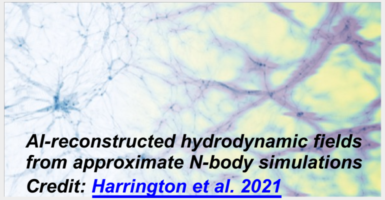
A changing computing landscape challenges us to think differently about supporting the Office of Science workload

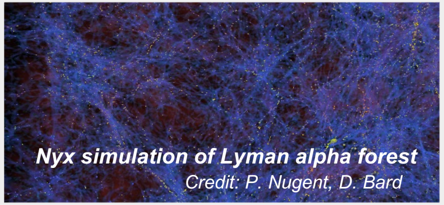
Growth of experimental and observational data and the need for interactive feedback through real-time data analysis and simulation and modeling

The proliferation of accelerators and new technologies



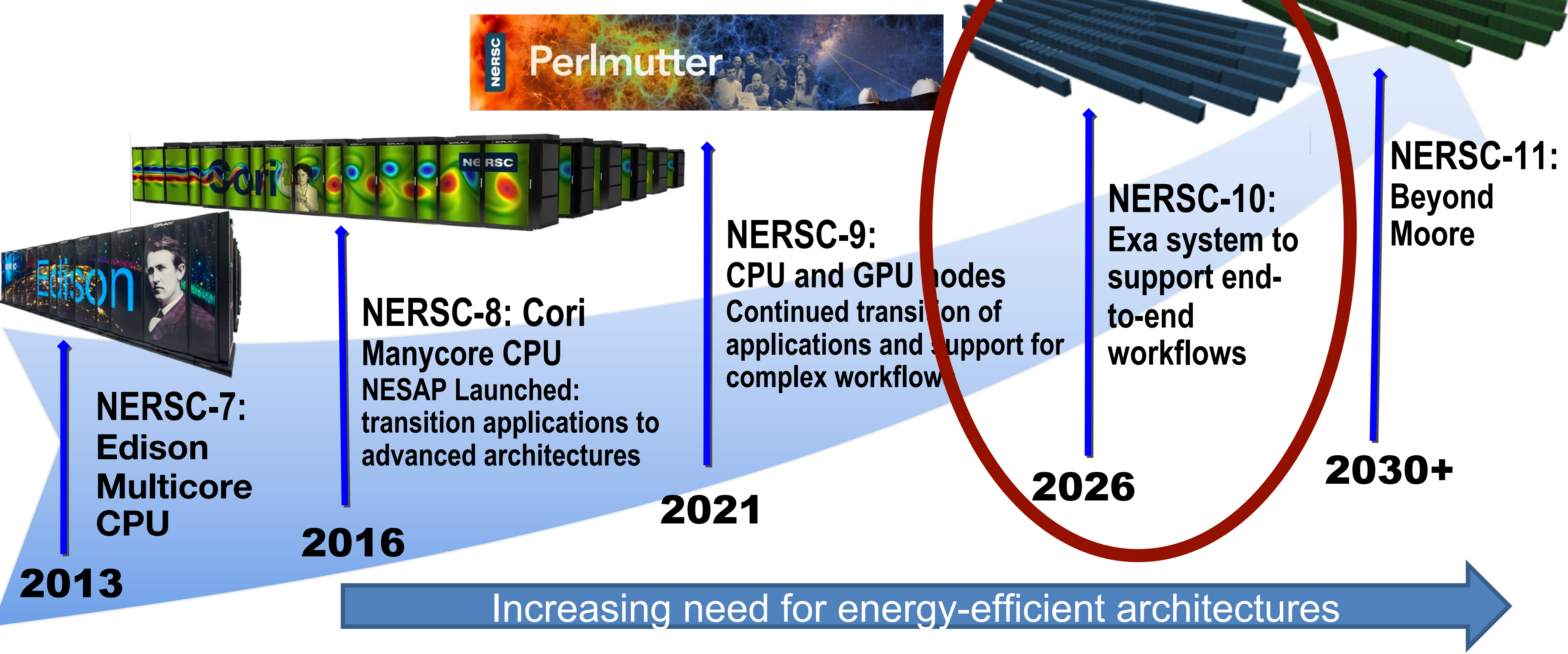
Use of advanced data analytics and AI in simulations as well as for integration of multimodal data sets


AI-reconstructed hydrodynamic fields from approximate N-body simulations
Credit: [Harrington et al. 2021]


Nyx simulation of Lyman alpha forest
Credit: P. Nugent, D. Bard

Credit: Katie Antypas (NERSC)

NERSC Systems Roadmap



Next Up: NERSC 10

Users require support for new paradigms for data analysis with **real-time interactive feedback between experiments and simulations**.

Users need the ability to search, analyze, reuse, and combine data from different sources into **large scale simulations and AI models.**

NERSC-10 Mission Need Statement:
The NERSC-10 system will ***accelerate end-to-end*** *DOE SC* ***workflows*** *and enable new modes of scientific discovery through the integration of experiment, data analysis, and simulation.*

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

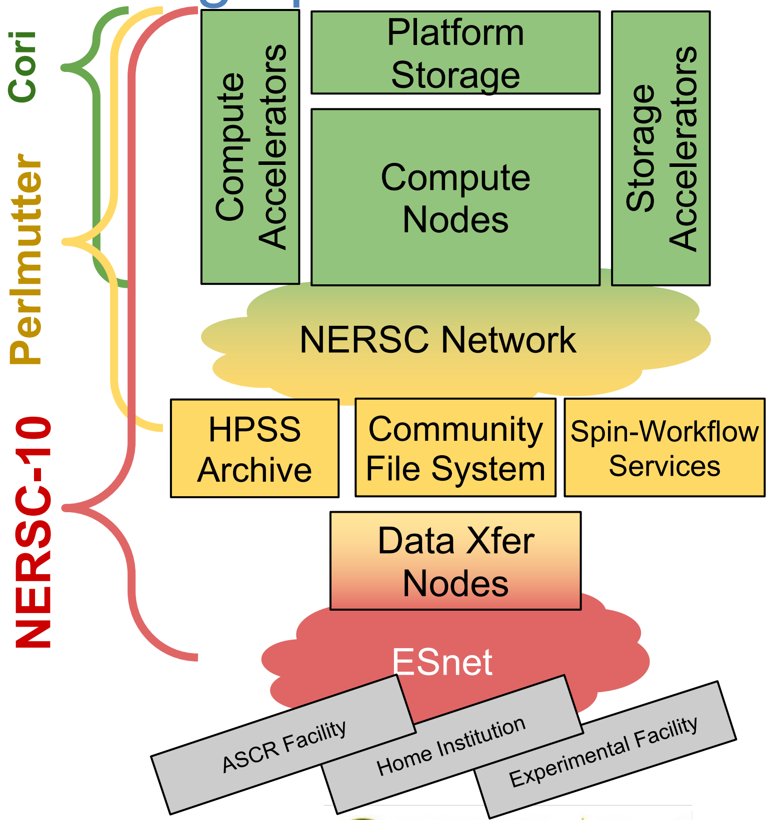
NERSC-10 Architecture: Designed to support complex simulation and data analysis workflows at high performance

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

Draft RFP released April 17th!
https://www.nersc.gov/systems/nersc-10/draft-tech-req/



Credit: Nick Wright, Hai Ah Nam (NERSC)

Summary

- The DOE runs a unique set of user facilities and national labs. When facilities can be used together we amplify the impact of science
- Experiments increasingly need supercomputing-scale resources for data analysis, simulation and digital twins
- NERSC is pioneering new modes of access to our systems to support experimental science
    - In addition to large-scale simulations, we can now support urgent workflows from experimental and observational facilities around the world
- The vision of the Integrated Research Infrastructure is to connect all DOE resources to “radically accelerate discovery and innovation”
- The HPC center workload is evolving quickly – combination of data analysis, AI and simulation, often in real time
    - N10 is designed to support complex workflows
- There are still many challenges – this is an exciting area of development in HPC!

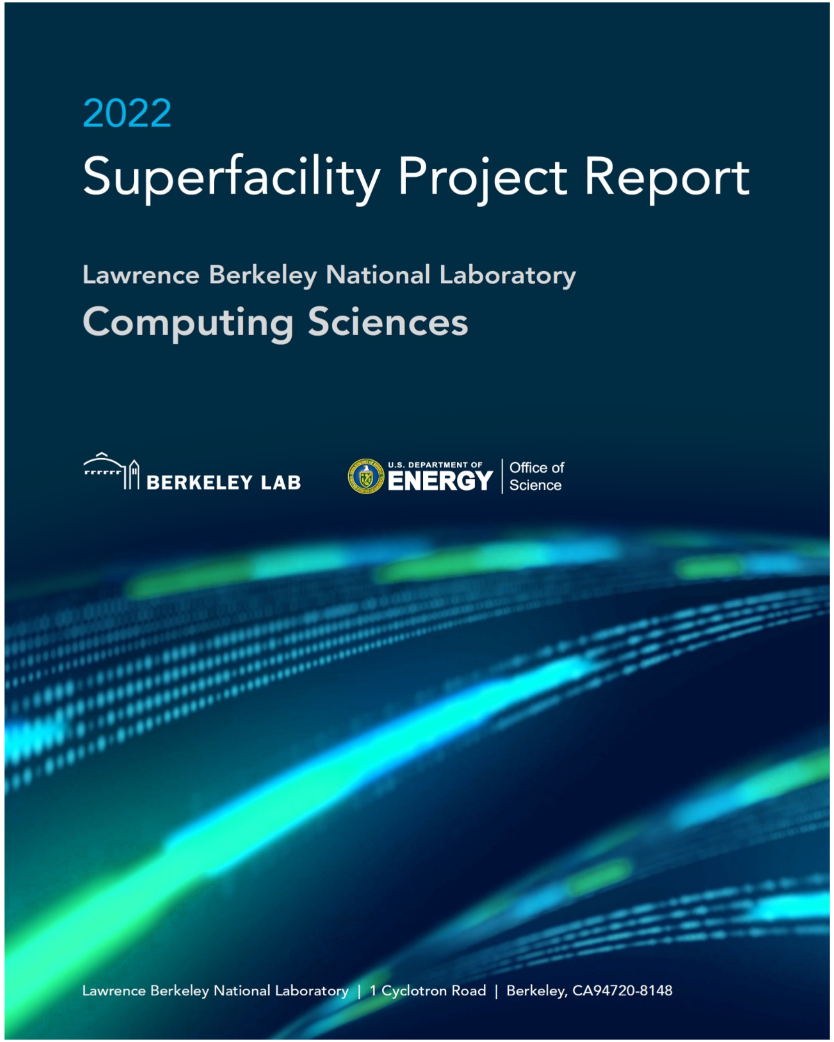
The Superfacility Project Report is now available and summarizes the work done, future priorities and lessons learned.

Thanks to everyone who contributed to it!

Debbie Bard, Cory Snavely, Lisa Gerhardt, Jason Lee, Becci Totzke, Katie Antypas, William Arndt, Johannes Blaschke, Suren Byna, Ravi Cheema, Shreyas Cholia, Mark Day, Bjoern Enders, Aditi Gaur, Annette Greiner, Taylor Groves, Mariam Kiran, Quincey Koziol, Tom Lehman, Kelly Rowland, Chris Samuel, Ashwin Selvarajan, Alex Sim, David Skinner, Laurie Stephey, Rollin Thomas, Gabor Torok

https://www.osti.gov/biblio/1875256
or google "superfacility project report"

AMCR
SciData


2022
Superfacility Project Report
Lawrence Berkeley National Laboratory
Computing Sciences
BERKELEY LAB
U.S. DEPARTMENT OF ENERGY | Office of Science
Lawrence Berkeley National Laboratory | 1 Cyclotron Road | Berkeley, CA94720-8148

Thanks!

NERSC