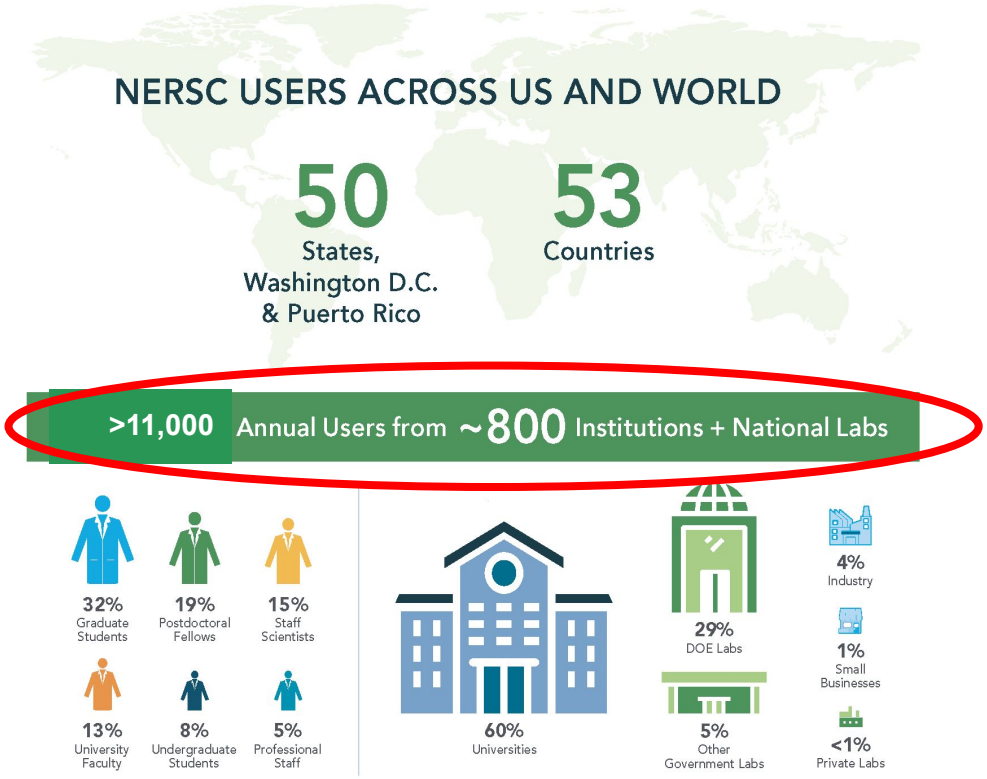
Architecting the NERSC "Doudna" system for complex workflows

Debbie Bard
Dept Head, Science Engagement and Workflows
Chair, Integrated Research Infrastructure Leadership Group

As the Mission HPC Center, NERSC is highly connected to the Office of Science

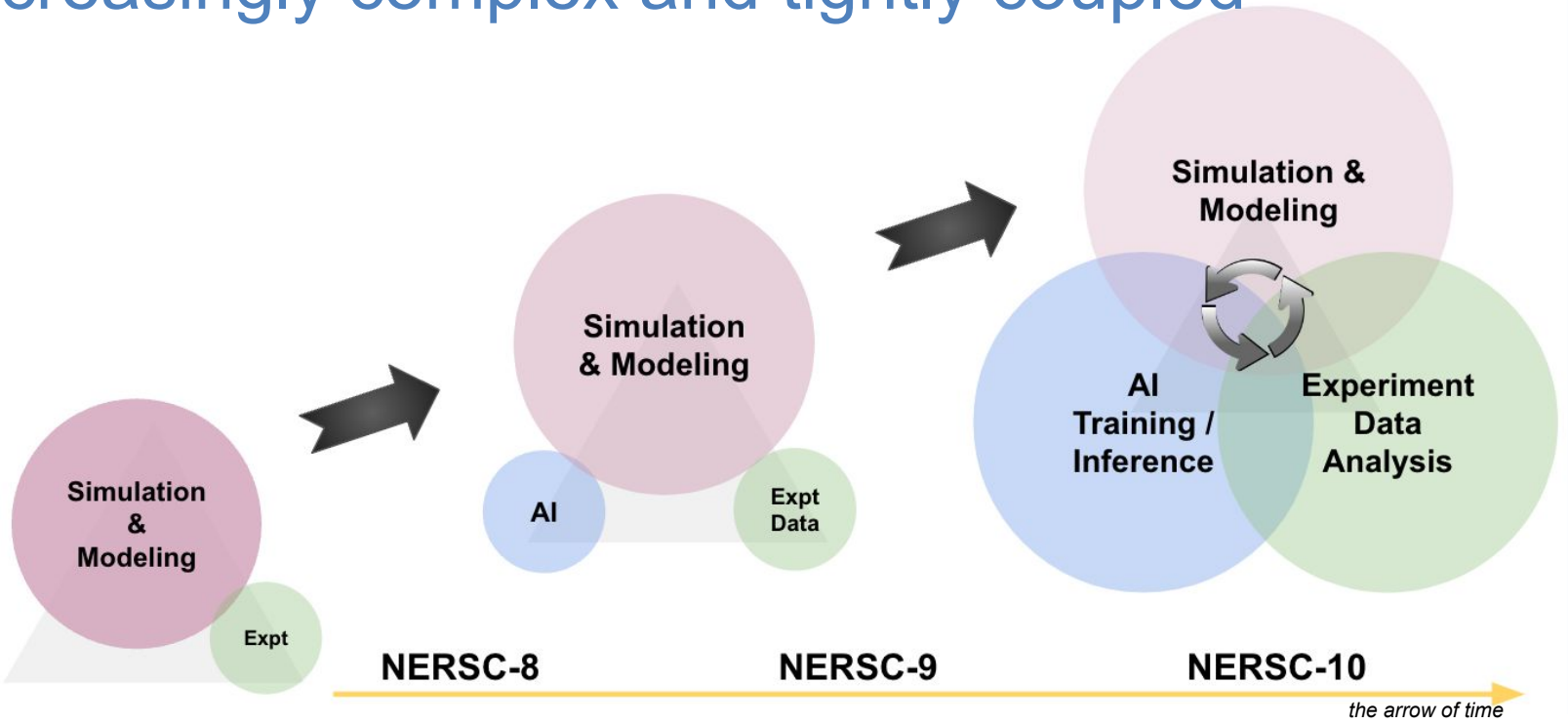


Our deep engagements with our user community give us a broad view of where the scientific computing is at and where it is heading.

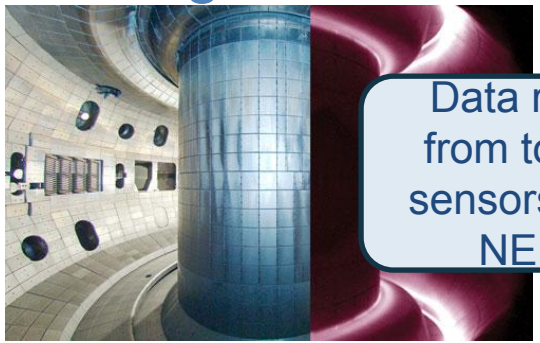
BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

We are seeing that science workflows are increasingly complex and tightly coupled

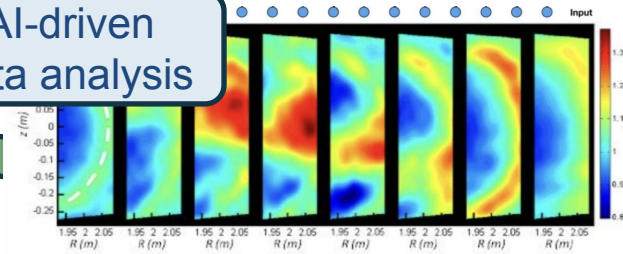


Science requires more integration across DOE facilities. DIII-D uses time-sensitive computing deeply embedded in an integrated framework

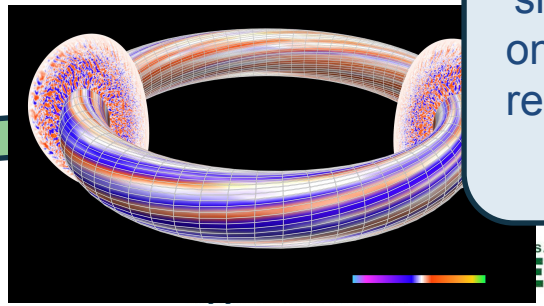


Data readout from tokamak sensors sent to NERSC

AI-driven data analysis

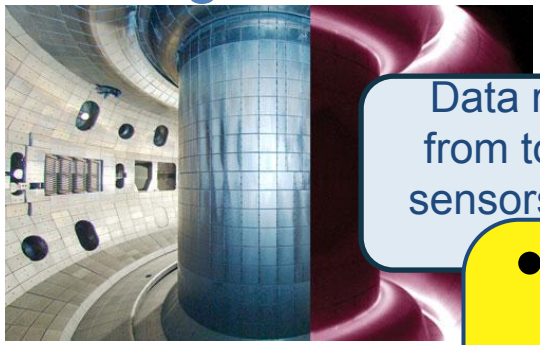


Digital Twin simulation based on analyzed data: recommends new experiment parameters



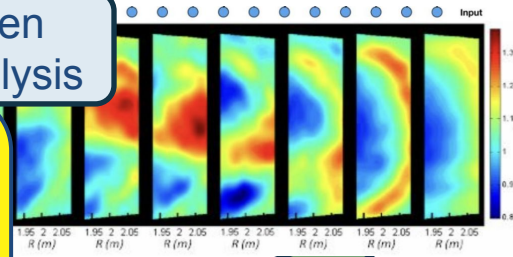
Feedback to scientist in minutes

Science requires more integration across DOE facilities. DIII-D uses time-sensitive computing deeply embedded in an integrated framework



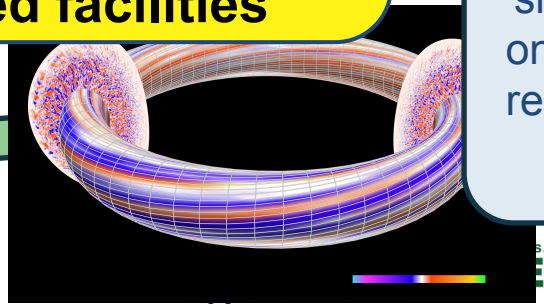
Data readout from tokamak sensors sent to

AI-driven data analysis



- **Automated, time-sensitive workflow**
- **High resolution modelling**
- **AI-based analysis**
- **Tightly connected facilities**

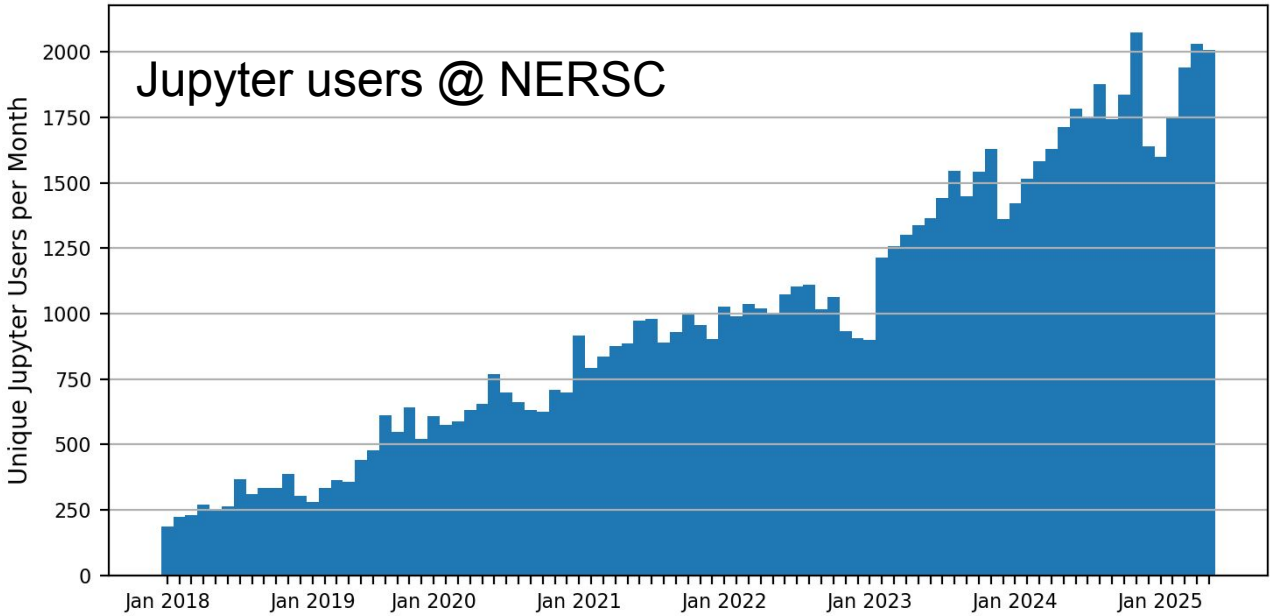
Feedback to scientist in minutes



Digital Twin simulation based on analyzed data: recommends new experiment parameters

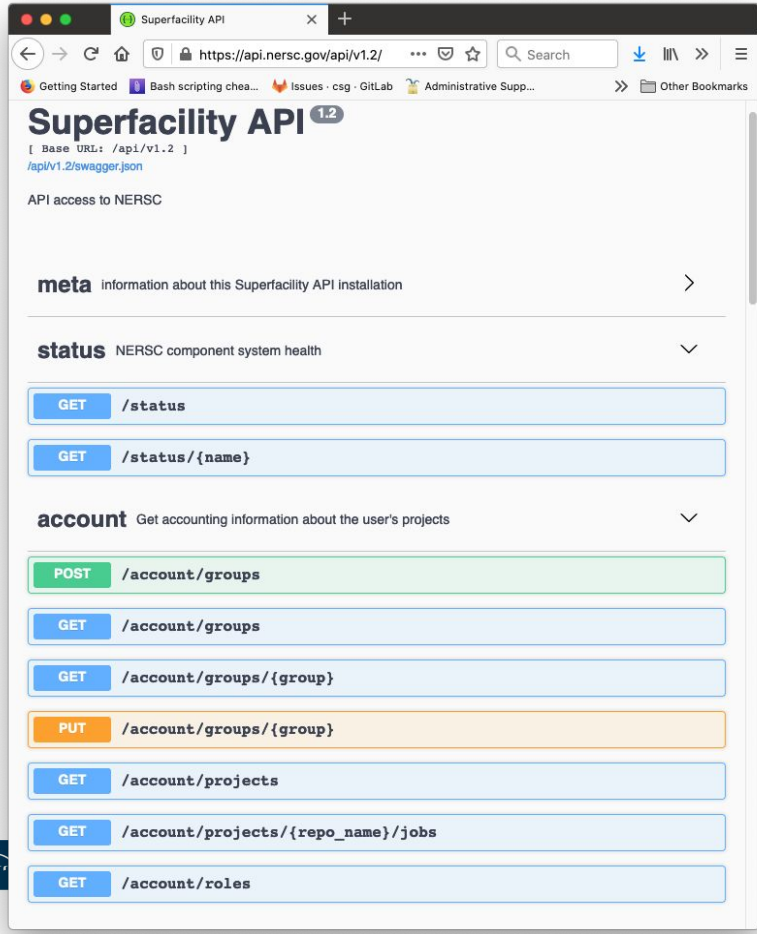
Users are interacting with our systems in new ways

- > 2.5k **Jupyter** users



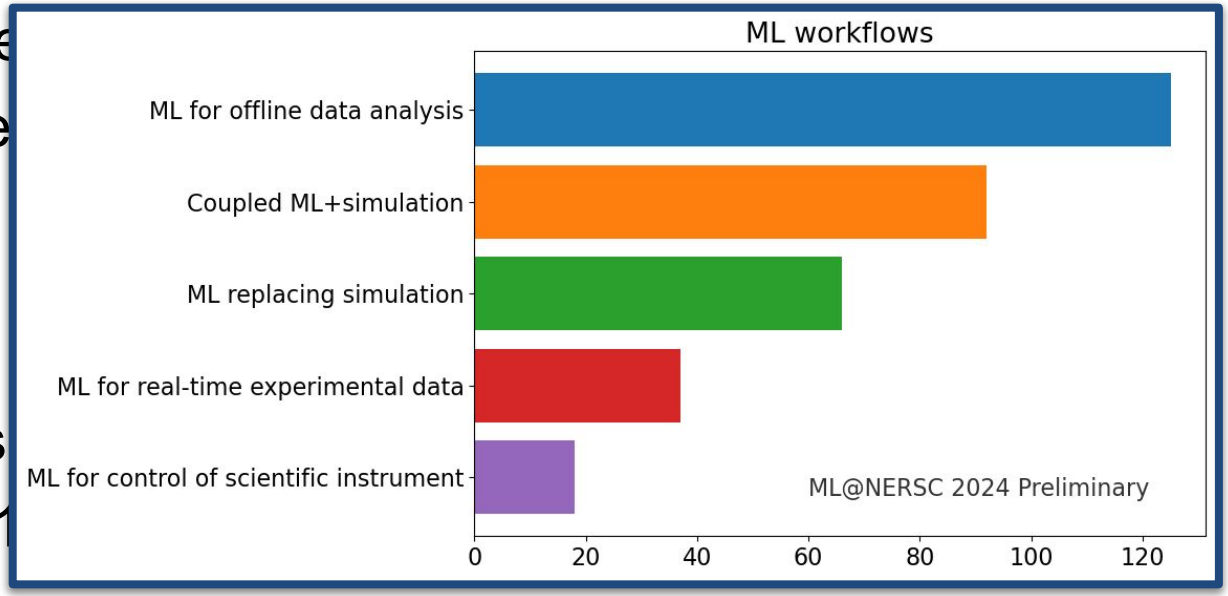
Users are interacting with our systems in new ways

- > 2.5k **Jupyter** users
- > 4.2k **Python** users - majority of active users
- Perlmutter's Top 500 result run in Shifter **container**
- **Federated Identity** used by >1000 users
- Superfacility **API**: 1 request logged every 2 sec

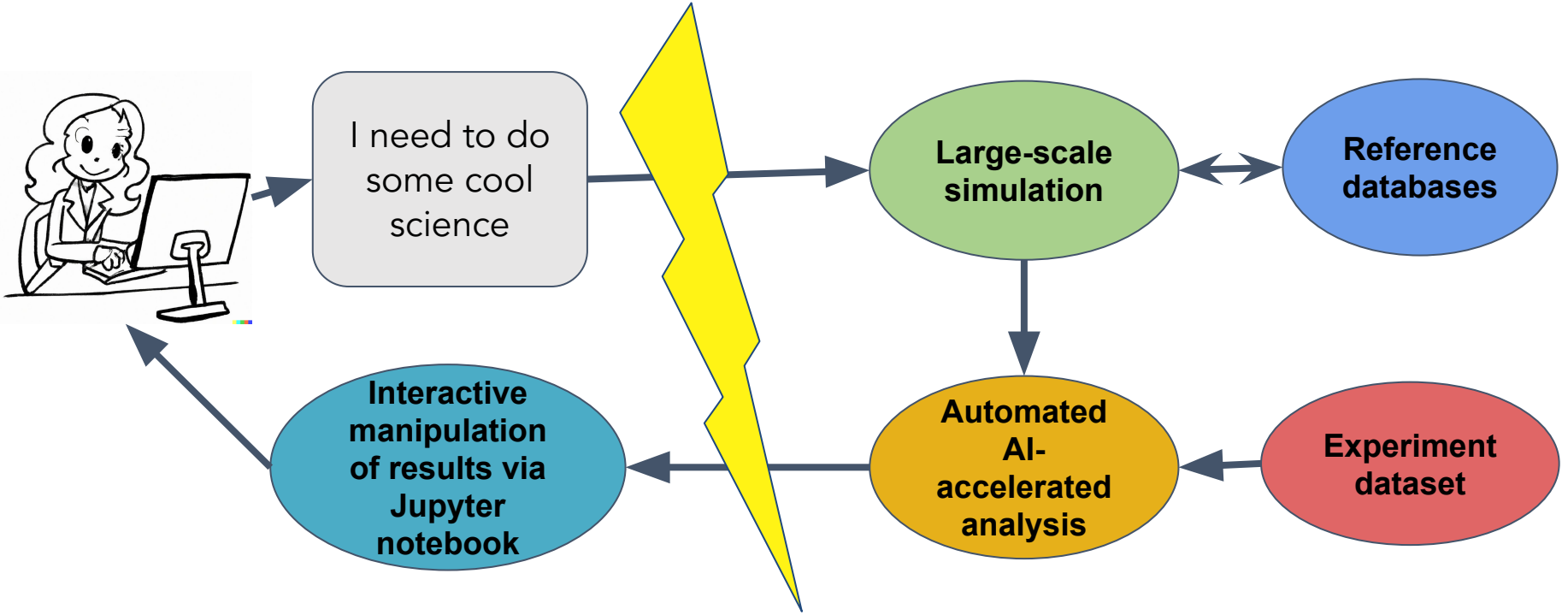


Users are interacting with our systems in new ways

- > 2.5k **Jupyter** use
- > 4.2k **Python** use
  active users
- NERSC's Top 500
  Shifter **container**
- **Federated IAM** us
- Superfacility **API**: 1
  every 2 sec
- >20x increase in **AI users** in 5 years



HPC centers need to provide the hardware, software, policies and training to support more complex workflows



HPC centers need to provide the hardware, software, policies and training to support more complex workflows

I need to do some cool science

magic happens

Large-scale

Reference bases

Key challenge: need to present these resources to scientists in a way that allows them to be productive

Interactive manipulation of results via Jupyter notebook

Automated AI-accelerated analysis

Experiment dataset

The technology landscape has changed a lot since we last bought a supercomputer

The NERSC-10 system market survey was broader than any we've done before:

- Hyperscalers
- AI accelerators
- Quantum computing
- Cloud technology
- AI-optimized storage vendors
- Specialised networks

NERSC-9 vendor Market Survey

NERSC-10 vendor Market Survey

The changing technology and vendor landscape makes it harder to provide HPC for the DOE scientific mission

Some science teams are ready to go! 🚀

Some science teams cannot respond to the pace of change

- don’t have specialist knowledge to adapt to specialist hardware
- don’t have algorithms that can adapt to new hardware
- don’t have funding to do the work

We cannot stagnate technologically, but we must also provide HPC for our full user base.

HPC centers need to provide the software, training and policies to support the use of new technologies

I need to do some cool science, using the best hardware for the job…

HPC for mod/sim

Quantum computer

Large-scale databases

Interactive manipulation of results

AI accelerators

HPC centers need to provide the software, training and policies to support the use of new technologies

I need to do some cool science, using the best hardware for the job…

magic happens

HPC for mod/sim

databases

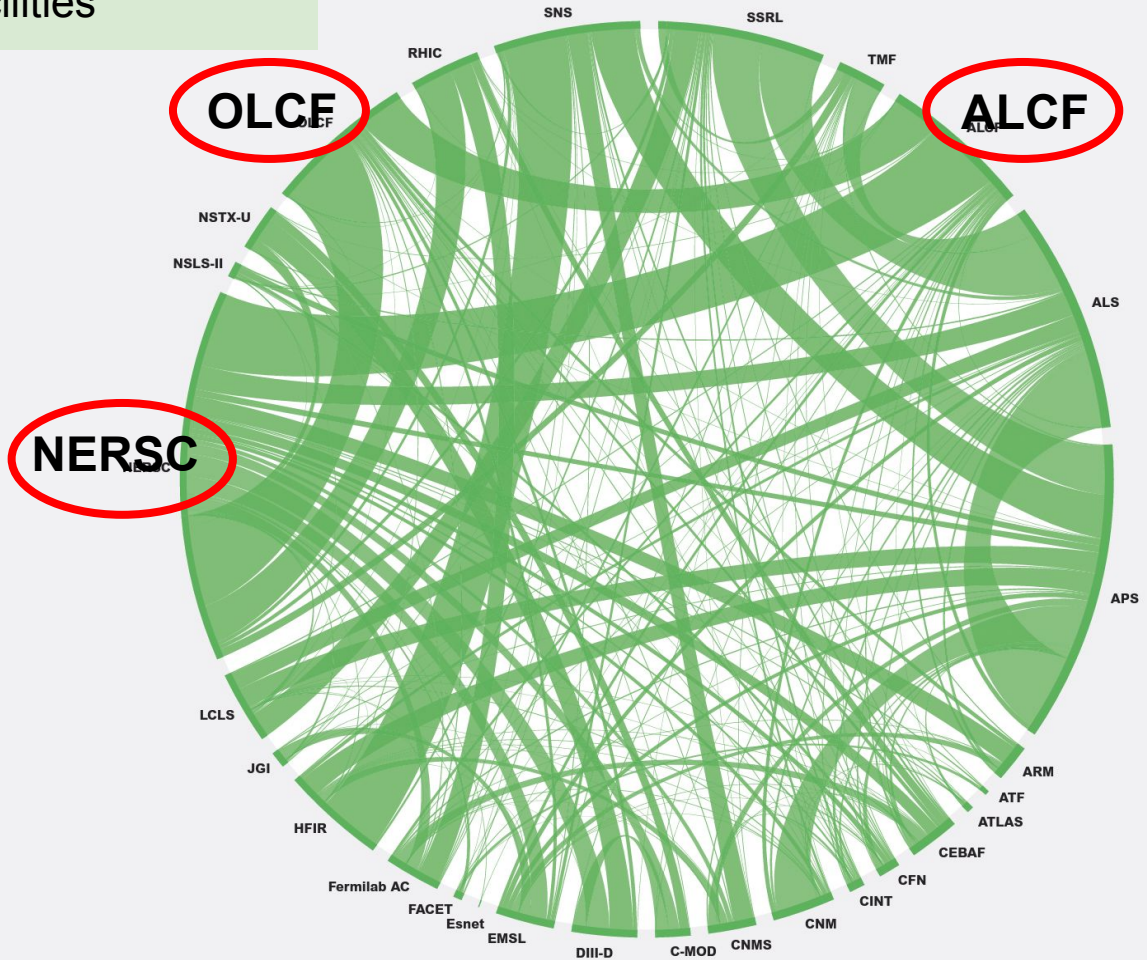
AI accelerators

Interactive manipulation of results

Key challenge: need to present these resources to scientists in a way that allows them to be productive

Each line represents one user using multiple facilities

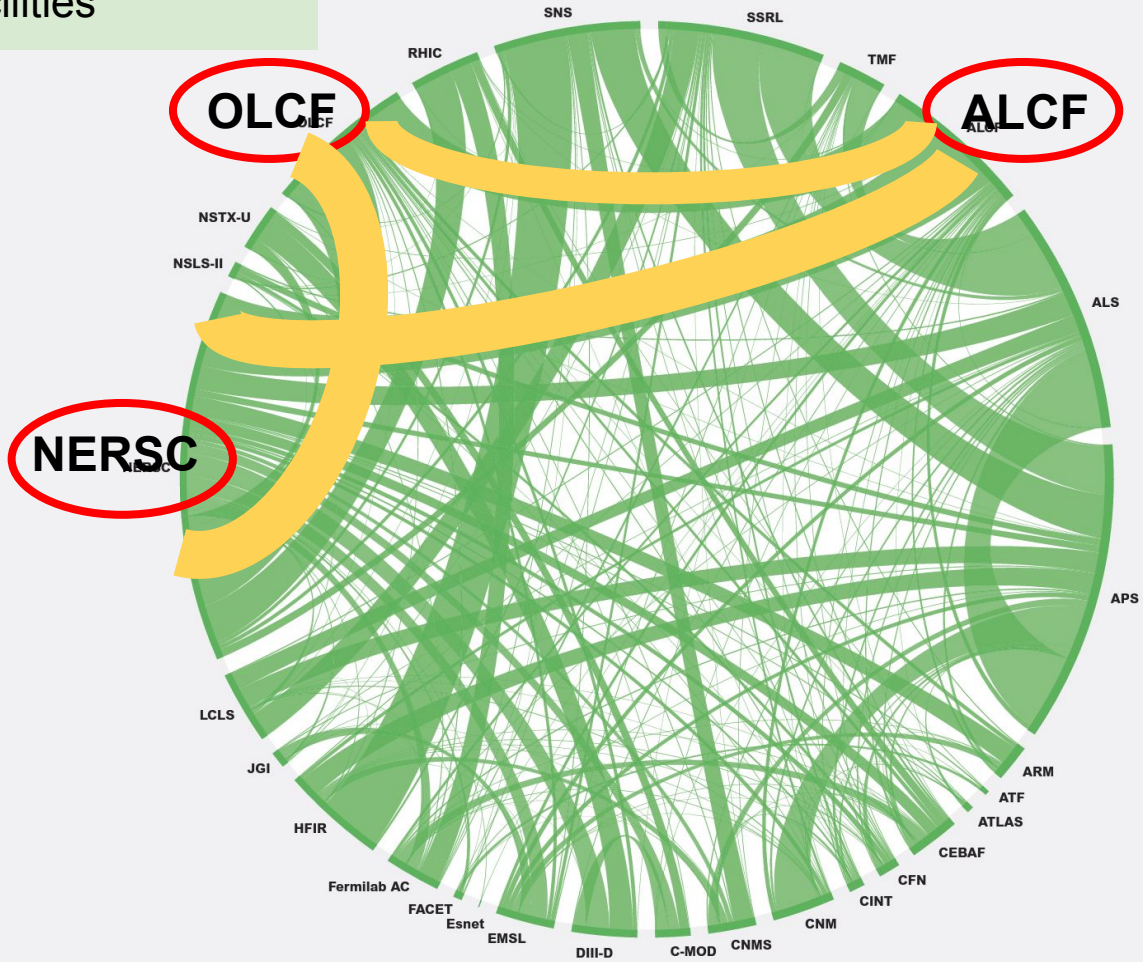
Users of one facility are increasingly users of multiple facilities.



Each line represents one user using multiple facilities

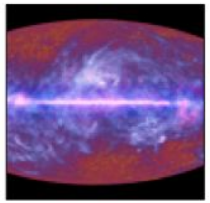
Users of one facility are increasingly users of multiple facilities.

Scientists don't just use one ASCR facility for their computing! Workflows span multiple computing centers.



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

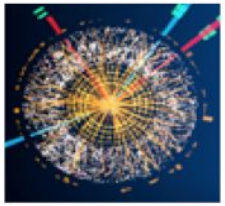
Palomar Transient Factory
Supernova



Planck Satellite
Cosmic Microwave Background Radiation

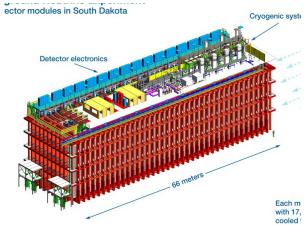
STAR

Star
Particle Physics

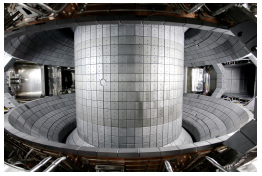


Atlas
Large Hadron Collider

APS



Dune

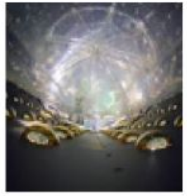


KStar

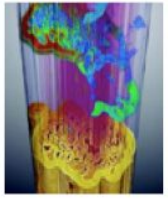
ARM

GLUEX

Katrin



Dayabay
Neutrinos

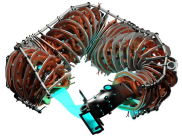


ALS
Light Source

LCLS
Light Source

Joint Genome Institute
Bioinformatics

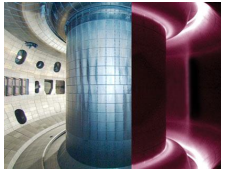
NSLS-II



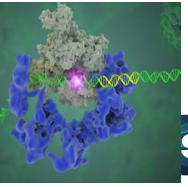
HSX

Majorana

AMERIFLUX

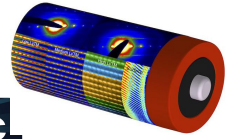


DIII-D

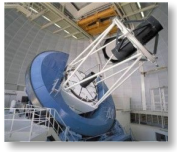


Cryo-EM

SC

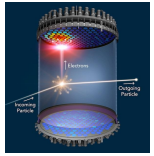


NCEM



DESI

LSST-DESC



LZ

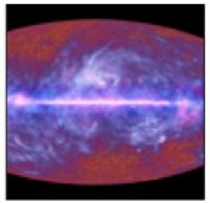
IceCube

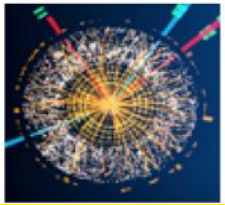
EXO

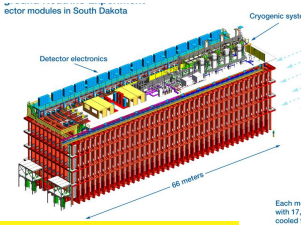
JBEI
Joint BioEnergy Institute

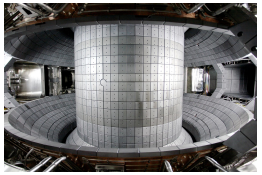
NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities

Palomar Transient Factory Supernova



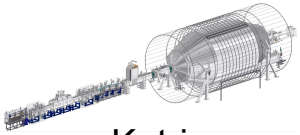






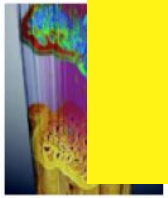
KStar

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

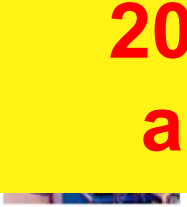


Katrin

Dayabay Neutrinos



ALS Light Source



LCLS Light Source

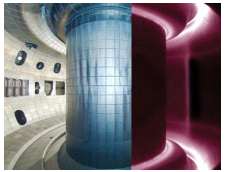
Joint Genome Institute Bioinformatics

NSLS-II

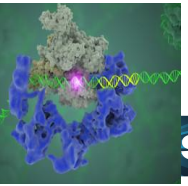
HSX

Majorana

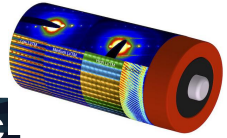
AMERIFLUX



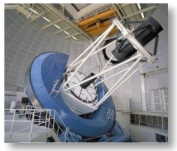
DIII-D



Cryo-EM

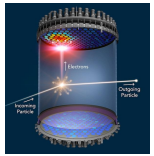


NCEM



DESI

LSST-DESC



LZ

IceCube

EXO

JBEI
Joint BioEnergy Institute

The 2024 ASCAC Facilities report emphasized the importance of an integrated facility ecosystem for DOE science

Recommendation 2: Science demands integration. We advocate viewing ASCR facilities not as isolated entities, but as integral components of a single, larger integrated computational *ecosystem*…, with a single governance model.

… Further, this integrated ecosystem is required for programs of other agencies, and industry. Its critical role in bolstering national scientific and technological capabilities, as well as its status as a model internationally, cannot be overstated.

https://science.osti.gov/-/media/ascr/ascac/pdf/reports/2024/FinalReport_May_2024_2370379.pdf

DOE needs to provide the software, policies and training to support integration across computing resources

I need to do some cool science

HPC facility 1

Large-scale simulation

Real-time feedback to scientist at a DOE user facility

Automated AI-accelerated analysis

HPC facility 2

Running Experiment

Experiment facility

DOE needs to provide the software, policies and training to support integration across computing resources

I need to do some cool science

magic happens

HPC facility 1

Large-scale

Key challenge: need to present these resources to scientists in a way that allows them to be productive

Automated AI-accelerated analysis

HPC facility 2

Running Experiment

Experiment facility

Real-time feedback to scientist at a DOE user facility

DOE has recognised this need, and is addressing some of the challenges via the **Integrated Research Infrastructure program (IRI)**

Advanced Networking

ESnet
ENERGY SCIENCES NETWORK

Edge Sensors

Local Campus Computing

Computing Testbeds

Cloud Computing

Researchers

Data Management

High Performance Data Facility

Data Repositories PuRE Data Assets

Software

Software and Applications

AI Tools Digital Twins

DOE established a FY24-25 Agency Priority Goal to stand up the **IRI Program.**

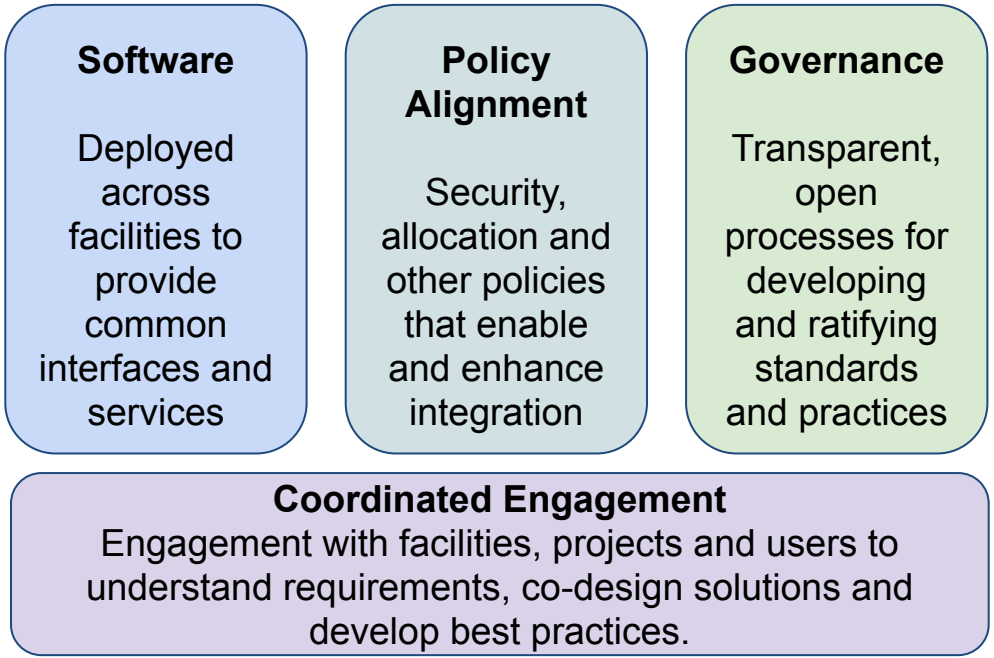
NERSC

IRI is creating the OS for the DOE Ecosystem

IRI is an "operating system" that layers on top of existing facilities.

IRI does not dictate the physical hardware to be deployed at each site, but facilities will need to make changes to support IRI software, policies and processes.

Facilities retain their unique mission and capabilities.



In FY25, we are taking the first steps towards a fully-featured IRI framework

IRI Program Area	FY25 Goal
IRI Allocations Program	• Develop **multi-year, multi-facility allocation programs**, for both R&D access and production systems
TRUSTID Design Patterns	• Define federated ID **design patterns** for trusted interoperable cross-facility workflows • Identify **policy changes** needed across the Office of Science
Interfaces	• Design a **minimal functional API** and deploy at multiple sites • Explore how to align **Jupyter** across sites
Software Deployment & Portability	• Align container deployment across sites
Scheduling/Preemption	• Align **real-time computing policies** across sites
Outreach and Engagement	• **Liaise with Pathfinder partners** to demonstrate cross-facility workflows using early IRI frameworks • Plan for **broader community engagement**

IRI has developed first prototype cross-facility API

- Unauthenticated status API specification and reference implementations: ✅
- IRI authentication middleware: coordinating with TRUSTID
- Authenticated API endpoints: Job submission, data movement, accounting queries…

```
[bcote@facility-api-prototype-vmw-01:~$ curl -k https://facility-api-prototype-vmw-01/resource
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   593  100   593    0     0  19006      0 --:--:-- --:--:-- --:--:-- 19129
{
  "id": "polaris_id",
  "name": "Polaris",
  "short_name": "polaris",
  "description": "ALCF HPE system with 560 nodes and a peak performance of 34 petaflops.",
  "last_modified": "2025-03-04T09:00:00Z",
  "links": [
    {
      "rel": "self",
      "href": "https://facility-api-prototype-vmw-01/resources/polaris_id"
    },
    {
      "rel": "dependsOn",
      "href": "https://facility-api-prototype-vmw-01/resources/eagle_id"
    },
    {
      "rel": "dependsOn",
      "href": "https://facility-api-prototype-vmw-01/resources/infiniband_id"
    },
    {
      "rel": "hasLastEvent",
      "href": "https://facility-api-prototype-vmw-01/events/event_4_id"
    }
  ],
  "type": "compute",
  "group": "computes",
  "status": "up"
}
```

IRI has developed first prototype cross-facility API

- Unauthe
  specifica
  impleme
- IRI auth
  coordina
- Authenti
  Job subi
  moveme
  queries…

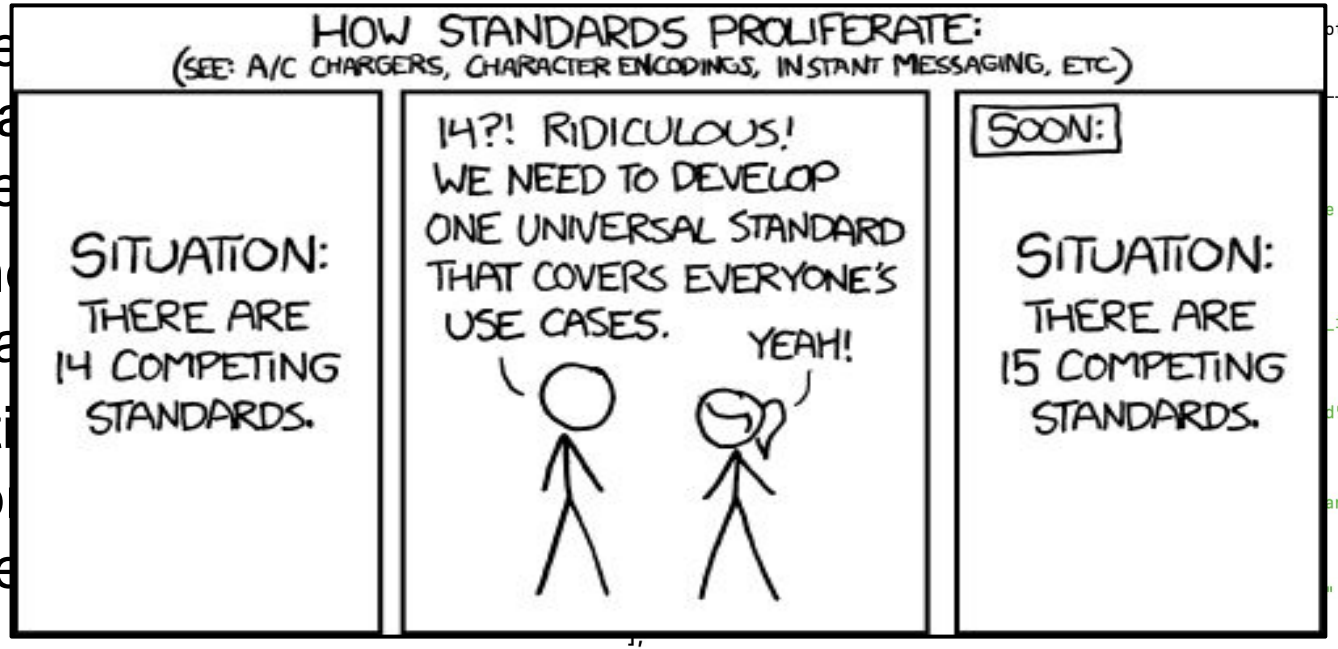


image credit: XKCD.com

NERSC system design is responding to the DOE Mission

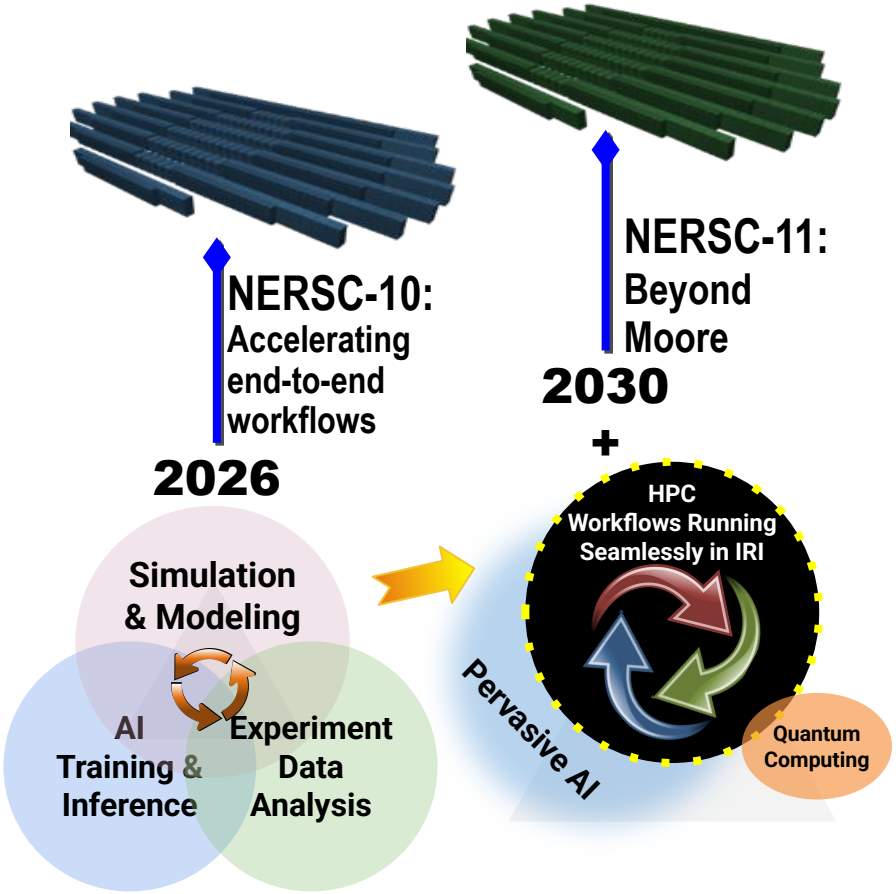
Scientists adopting changes in the technology landscape
Mixed precision & AI-based surrogates
Adoption of cloud technologies

Scientists leveraging **new hardware to do new science**
Quantum computing

Continued and increased demand for **higher resolution and larger domain mod/sim,** incorporating additional physics

Increasing automation
Including AI-driven automation
APIs everywhere

Increased connectivity
between experiment and compute sites, driven by IRI



NERSC-10: A Supercomputer for Complex, Integrated Workflows

Users require an ***integrated ecosystem*** that supports 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 (2021): 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.

NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science

The NERSC-10 system was announced in May, named in honor of **Jennifer Doudna**, the Berkeley Lab-based biochemist who was awarded the 2020 Nobel Prize for Chemistry for her work on the gene-editing technology CRISPR..

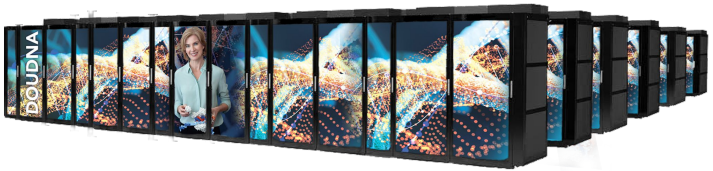
DOE's Commitment to Advancing American Leadership in Science, AI and HPC

"The *Doudna* system represents DOE's commitment to advancing American leadership in science, AI, and high-performance computing," said **U.S. Secretary of Energy Chris Wright**. "It will be a powerhouse for rapid innovation that will transform our efforts to develop abundant, affordable energy supplies and advance breakthroughs in quantum computing. AI is the Manhattan Project of our time, and *Doudna* will help ensure America's scientists have the tools they need to win the global race for AI dominance."

"At Dell Technologies, we are empowering researchers worldwide by seamlessly integrating simulation, data, and AI to address the world's most complex challenges," said **Michael Dell, Chairman and CEO, Dell Technologies**. "Our collaboration with the Department of Energy on Doudna underscores a shared vision to redefine the limits of high-performance computing and drive innovation that accelerates human progress."

"*Doudna* is a time machine for science — compressing years of discovery into days," said **Jensen Huang, founder and CEO of NVIDIA**. "Built together with DOE and powered by NVIDIA's Vera Rubin platform, it will let scientists delve deeper and think bigger to seek the fundamental truths of the universe."

Doudna System Overview



Designed to provide > 10x performance over *Perlmutter* and support diverse and complex DOE SC workflows.

NVIDIA User Software

Vera-Rubin CPU-GPU

NVIDIA **Vera-Rubin Compute Node**

- Able to support GPU and CPU portions of complex workflows
- Integrated AI and Compute Optimized Partitions

High Speed Network

NVIDIA **Infiniband XDR (800 Gbps)**

- Quantum-3 Infiniband switches
- ConnectX-8 Infiniband NICs
- Unified Fabric Manager

External Connectivity

- Skyway-Next IB-to-Eth Gateway

2 Types of Storage

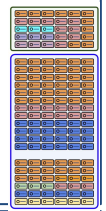
IBM **Platform Storage System (PSS)**
IBM Storage Scale System

VAST **Quality of Service Storage System (QSS)** on Dell PowerEdge Servers

Workflow Environment

Workflow Environment Nodes
Heterogeneous node-types
CPU-only, Vera-Rubin
Air-cooled, Water-cooled

Reconfigurable to support complex IRI workflows (e.g. batch, compile, Jupyter, cloud-native, data transfer, etc.)



Dell System Management Tools (OME, iDRAC, Omnia)

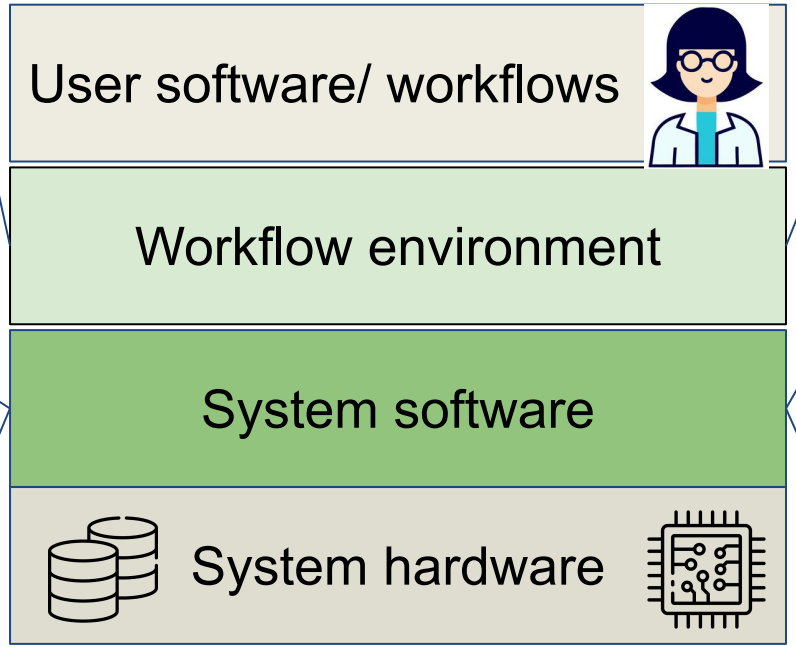
Dell ORv3 Direct liquid-cooled server technology & integrated rack scalable systems

Innovation in software is key to enabling complex workflows on *Doudna*

New capabilities: FaaS/serverless, specialized HW, AI deployment, data lifecycle, quantum…

Support usage of both ssh and Jupyter

Meet federal security requirements



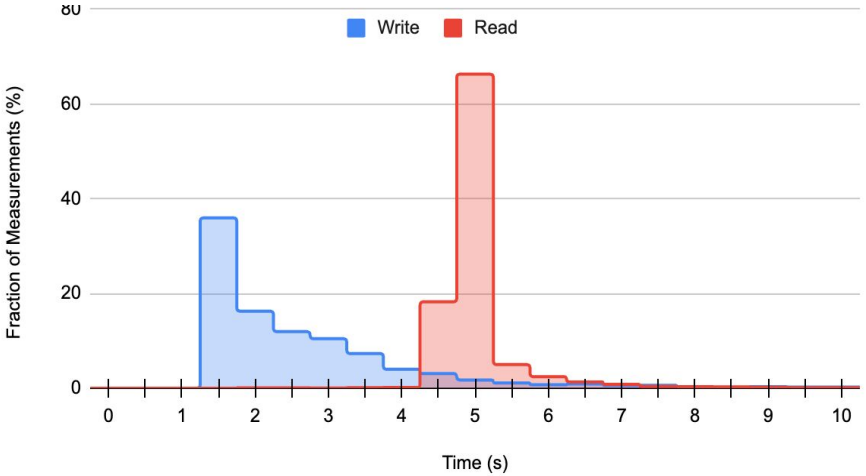
RESTful user-facing APIs to support automation

System-side APIs for workflow observability, administration and reconfigurability

Containerize the user environment

The NERSC workload requires capabilities that are hard to reconcile in a single file system on *Doudna*

IOR performance on Perlmutter



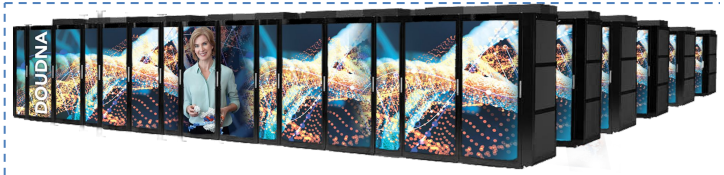
- 21% of all write tests took more than twice as long as the mode (1.5 sec)
- 2% of all write tests took at least **five times longer** than the mode

For instrument-driven and time-dependent workflows such variance could be catastrophic

- **Quality of Service Storage System (QSS)** will provide controllable, guaranteed IOPs / bandwidth to meet the needs of time-sensitive workflows
- Platform Storage System (PSS) is a more traditional FS that will meet the needs of much of the NERSC workload

Doudna Drives the Integrated Data Center Ecosystem

Scientific breakthroughs in AI and at Experimental Facilities enabled through high-speed access to DOE SC Community Data



> 10x performance over *Perlmutter* and supports diverse and complex DOE SC workflows

Workload Optimized Compute Capabilities

- Able to support GPU and CPU portions of complex workflows
- Integrated AI and Compute Optimized Partitions

2 Types of Storage Capabilities

Platform Storage System (PSS) For large-scale simulation & modeling

Quality of Service Storage System (QSS) for data intensive AI and Experiment

Workflow Environment

Workflow Environment Nodes

Reconfigurable to support complex IRI workflows (e.g. batch, compile, Jupyter, cloud-native, data transfer, etc.) with integrated Spin capabilities.

200 GB/s

HPSS Data Archive (375 PB)
Demand is growing by 30% each year
Data can be reused & leveraged for AI training

> 800 GB/s

Community File System (150 PB)
Allows sharing data between users, systems and the 'outside'

3.25 TB/s (26 Tbps)

NERSC Data Center Network
Ethernet LAN

> 10 GB/s

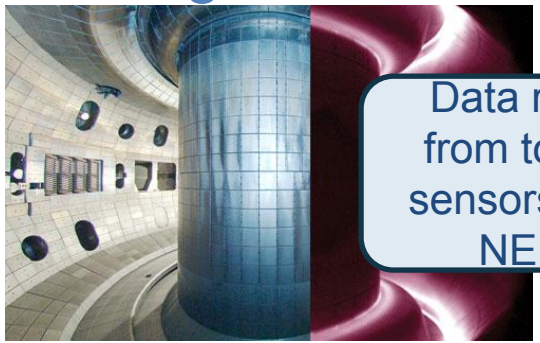
/home (400 TB)
Permanent storage and backup for important files

Data Transfer Nodes
Transfers data between NERSC storage resources and externally to resources at other sites

ESnet ENERGY SCIENCES NETWORK
2 x 400 Gb/s
2 x 100 Gb/s

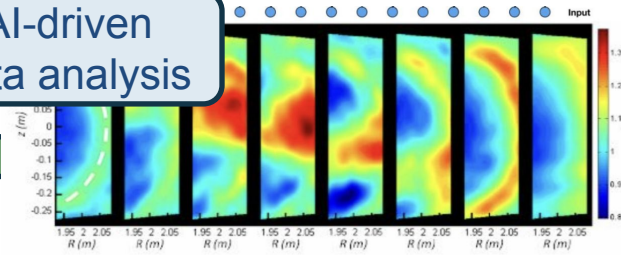
Experimental Facility | **ASCR Facility** | **Home Institution** | **Cloud** | **Edge**

Science requires more integration across DOE facilities. DIII-D uses time-sensitive computing deeply embedded in an integrated framework

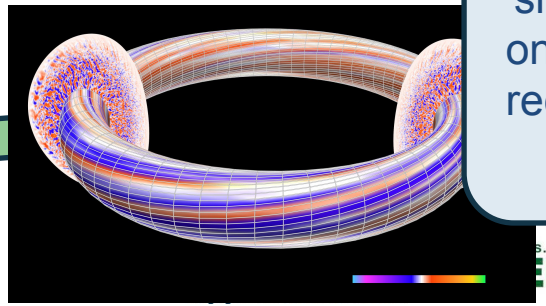


Data readout from tokamak sensors sent to NERSC

AI-driven data analysis

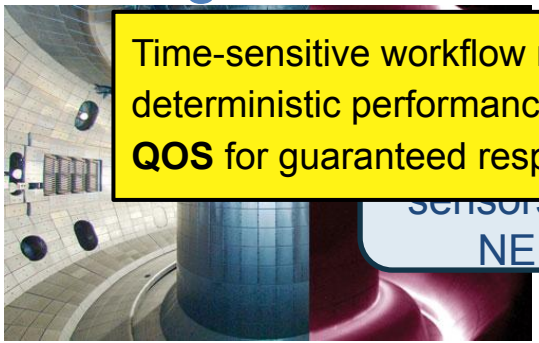


Digital Twin simulation based on analyzed data: recommends new experiment parameters



Feedback to scientist in minutes

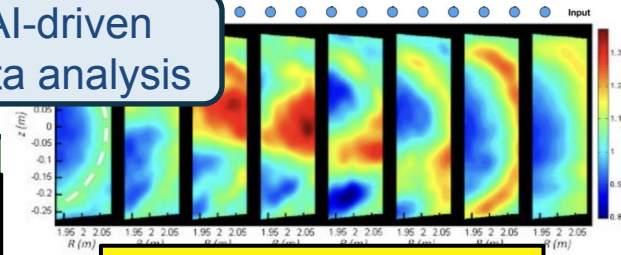
Science requires more integration across DOE facilities. DIII-D uses time-sensitive computing deeply embedded in an integrated framework



Time-sensitive workflow requires **QSS** for deterministic performance and **network QOS** for guaranteed response in O(min)

sensors sent to NERSC

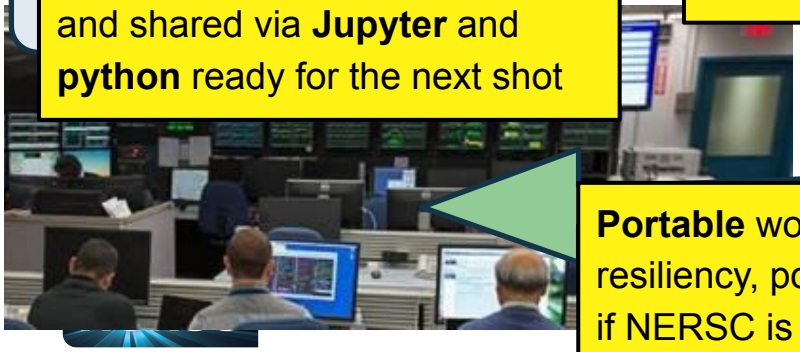
AI-driven data analysis



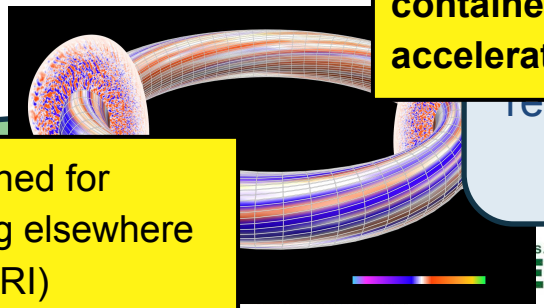
Data movement and compute progress tracked using **APIs** by automated workflow orchestrator

Feedback to

Results synthesized, displayed and shared via **Jupyter** and **python** ready for the next shot



Large-scale analysis and simulation use **containerized** apps and **accelerated** nodes.



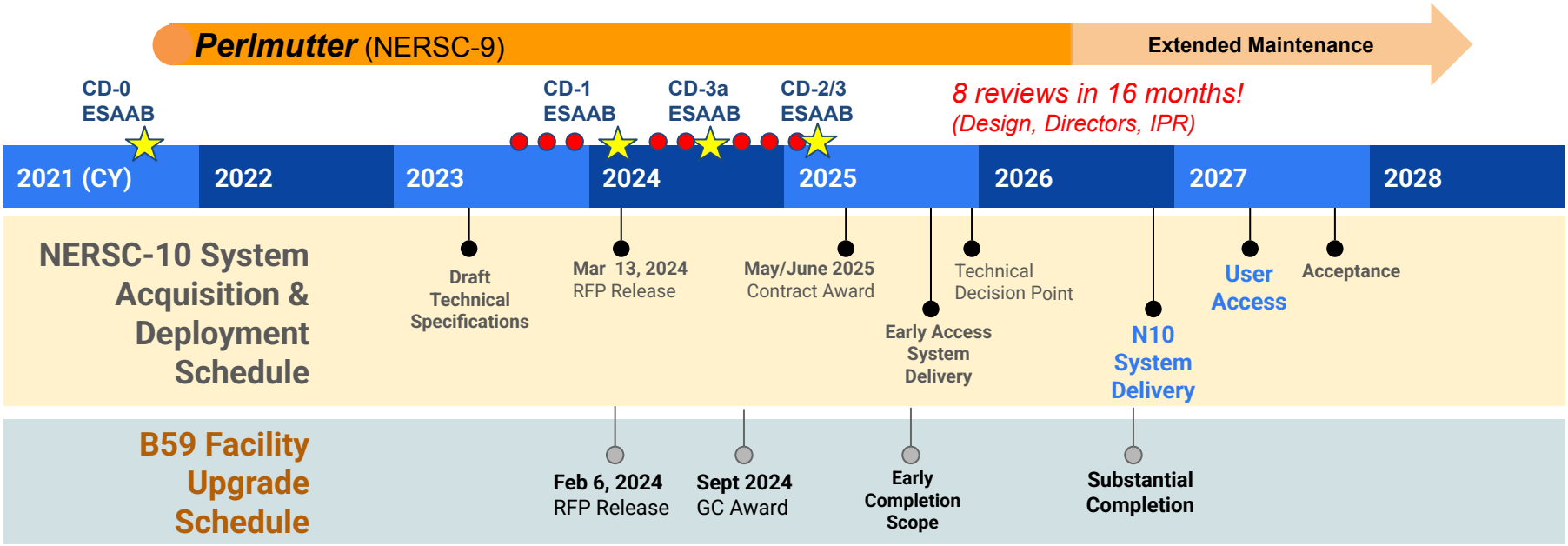
d
a:
recommends new experiment parameters

Portable workflows designed for resiliency, possibly running elsewhere if NERSC is unavailable (IRI)

S. DEPARTMENT OF ENERGY | Office of Science

Doudna Major Milestones

***Doudna* must provide user access in 2027** to maximize user productivity and ensure sufficient time to migrate 11,000+ DOE SC users before *Perlmutter* is decommissioned.



The *Doudna* system will accelerate end-to-end DOE SC workflows and enable new modes of scientific discovery through the integration of simulation, data analysis and experiment.

Our technology choices for Doudna are informed by the work we've done over the past 5 years to develop, operationalize and support Perlmutter and our users - including lessons learned from the Superfacility project and IRI.

- ***Doudna will deliver 10x Perlmutter performance on HPC workflows.***
- ***Doudna is designed to be IRI-ready.***
- ***GPU-enabled applications should have minimal issues in porting/running their applications.***
- ***Doudna is expected to be delivered in late 2026.***

Thanks!


DOUDNA
ACCELERATING
INNOVATION
NERSC
DELL Technologies
NVIDIA
BERKELEY LAB
U.S. DEPARTMENT of ENERGY