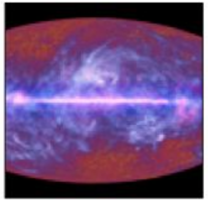
NERSC

NERSC-10 and the Integrated Research Infrastructure program

Debbie Bard
Data Department Head
22nd Feb 2024

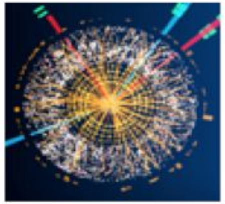
NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities

Palomar Transient Factory Supernova



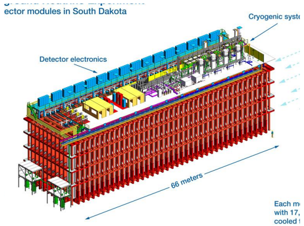
Planck Satellite Cosmic Microwave Background Radiation

Star Particle Physics

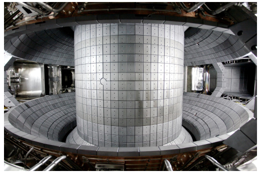


Atlas Large Hadron Collider

APS



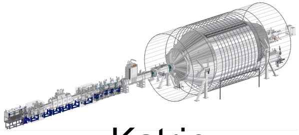
Dune



KStar

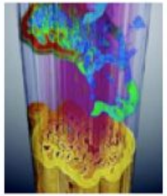
ARM

GlueX



Katrin

Dayabay Neutrinos

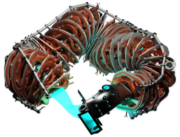


ALS Light Source

LCLS Light Source

Joint Genome Institute Bioinformatics

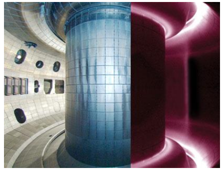
NSLS-II



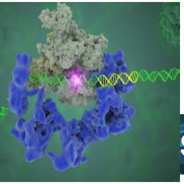
HSX

Majorana

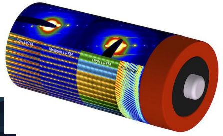
AmeriFlux



DIII-D



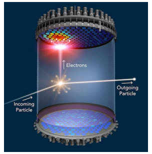
Cryo-EM



NCEM

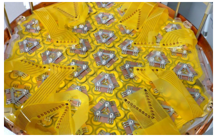
DESI

LSST-DESC



LZ

IceCube



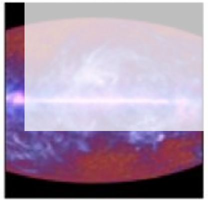
EXO

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

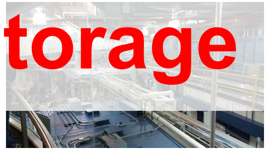
Palomar Transient Factory Supernova



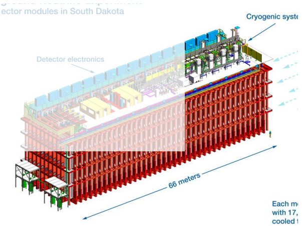
Planck Satellite Cosmic Microwave Background Radiation

Star Particle Physics

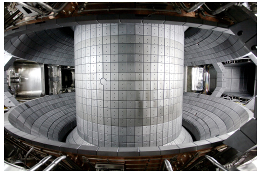
Atlas Large Hadron Collider



APS



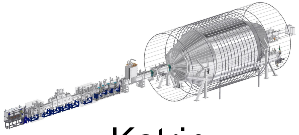
Dune



KStar

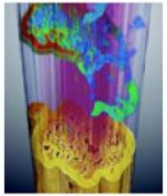
ARM

GlueX



Katrin

Dayabay Neutrinos

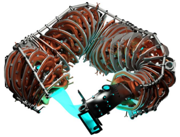


ALS Light Source

LCLS Light Source

Joint Genome Institute Bioinformatics

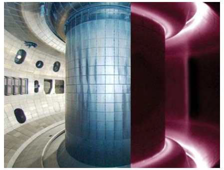
NSLS-II



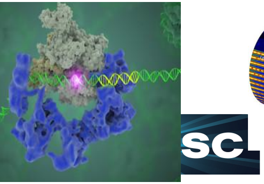
HSX

Majorana

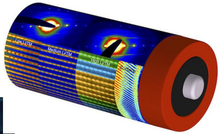
AmeriFlux



DIII-D



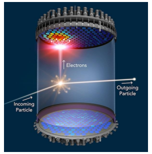
Cryo-EM



NCEM

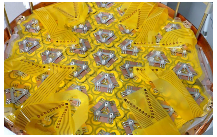
DESI

LSST-DESC



LZ

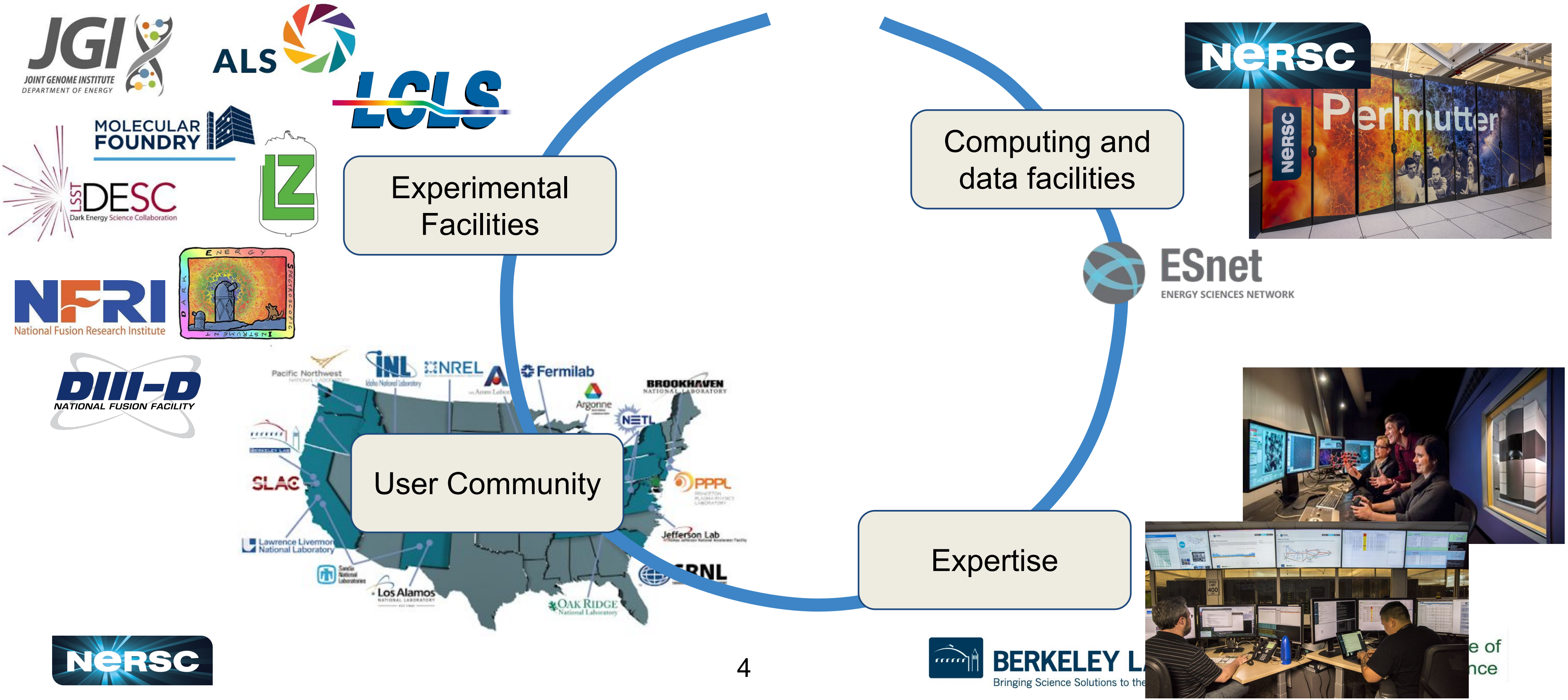
IceCube



EXO

JBEI
Joint BioEnergy Institute

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



Multiple science teams are using NERSC for superfacility-enabled science, in production

The Superfacility project (2019-2022) kick-started this work, building the base infrastructure and services. We now support multiple science teams using automated pipelines to analyze data from remote facilities at large scale.

- **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

Multiple science teams are using NERSC for superfacility-enabled science, in production

The Superfacility project (2019-2022) kick-started this work, building the base infrastructure and services. We now support multiple science teams using automated pipelines to analyze data from remote facilities at large scale.

- **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**

>20 science teams use the **realtime** qos to process urgent data

>40 projects use the NERSC **API**, ~19M logged requests since May 2022 = one request every 2 sec

>1500 unique **Jupyter** users per month, similar to number of users who ssh into our systems

>250 users, >85 projects use **Spin**

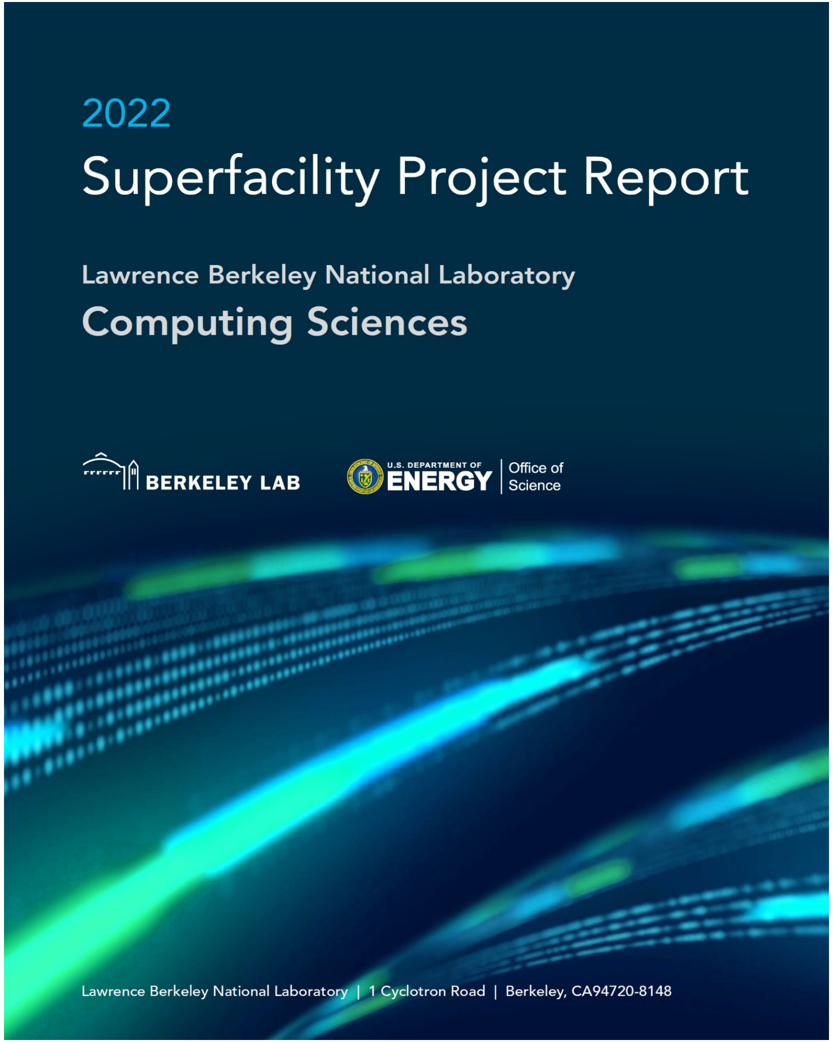
The Superfacility Project Report summarizes the work done during the project phase, future priorities and lessons learned.

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

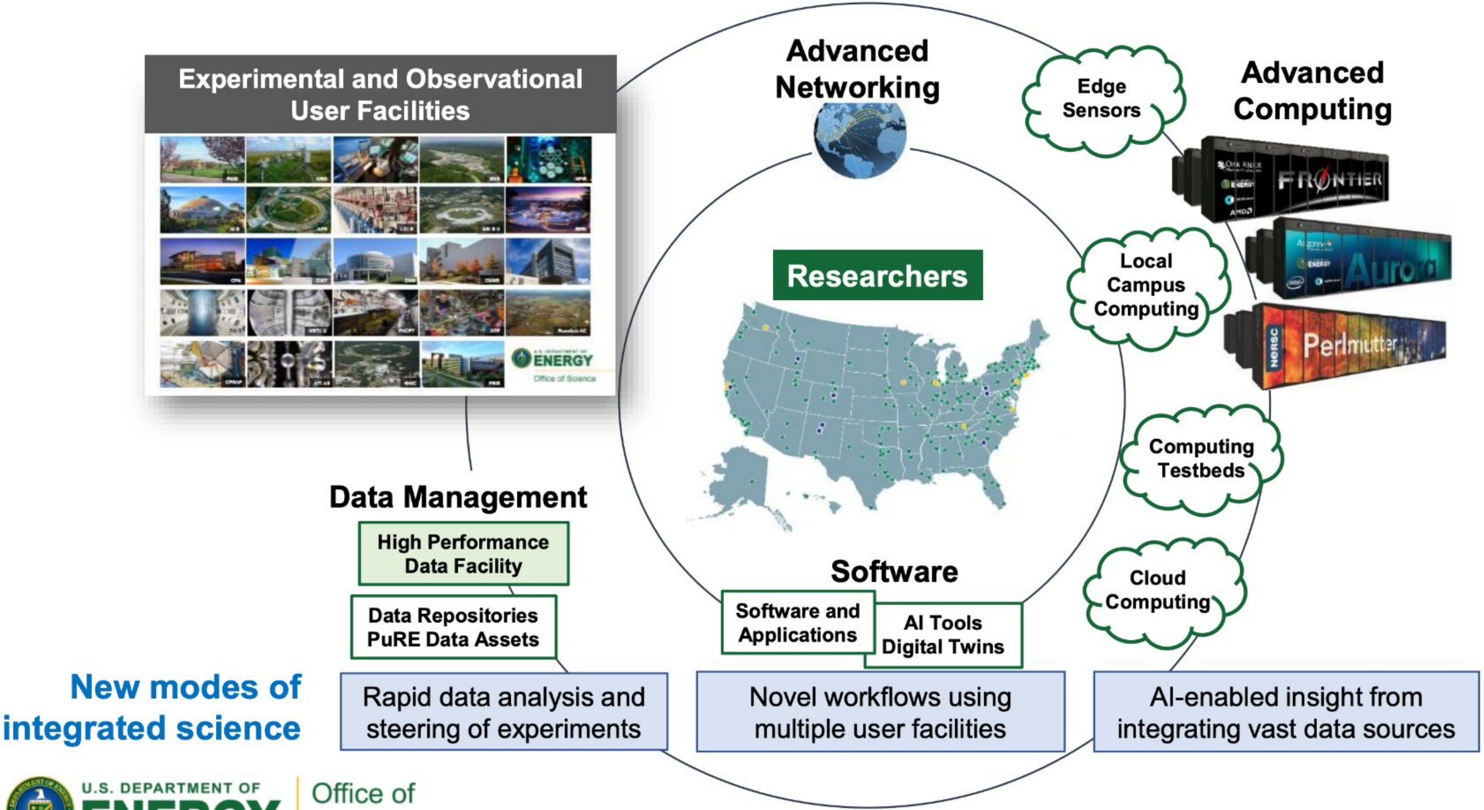
https://www.osti.gov/biblio/1875256
or search “superfacility project report”

AMCR
SciData



DOE's Integrated Research Infrastructure (IRI) Vision:

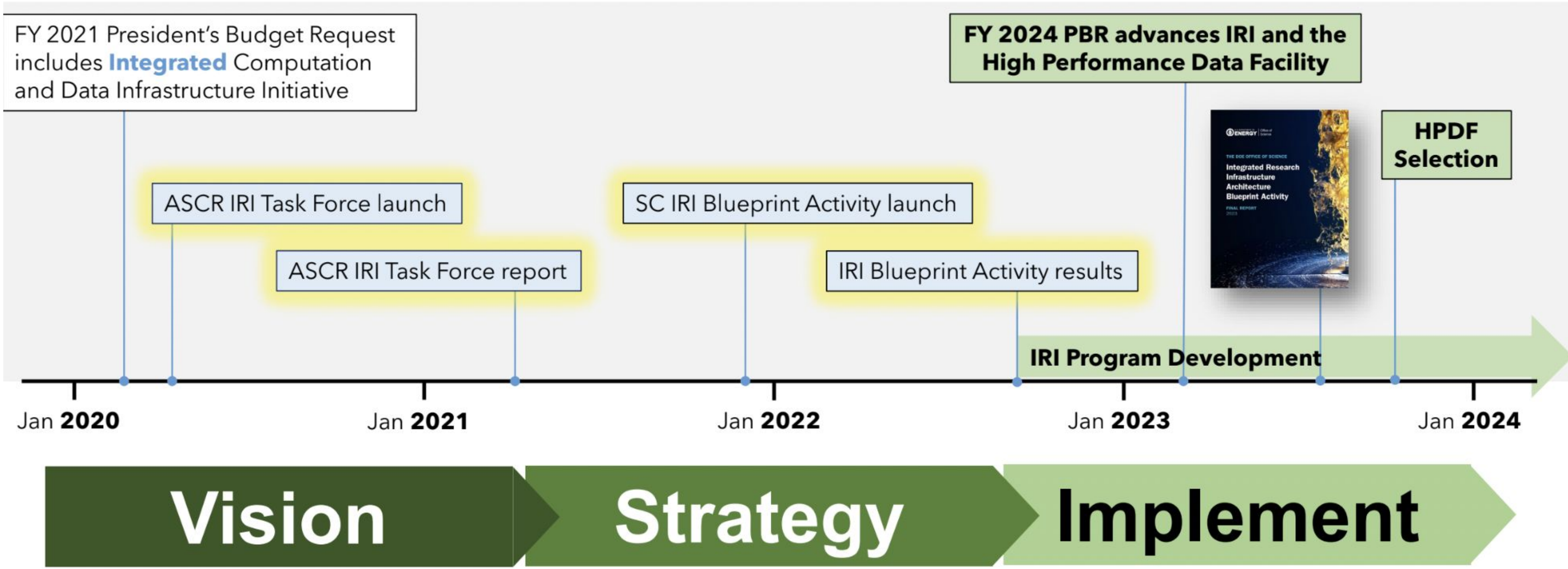
To empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation



U.S. DEPARTMENT OF ENERGY | Office of Science

Slide adapted from Ben Brown, ASCR

Timeline of IRI Program Development



Slide adapted from Ben Brown, ASCR

The 2022 IRI Architectural Blueprint Activity identified 6 key challenge areas and requirements from science teams

The IRI Framework comprises:

- 3 IRI Science Patterns represent integrated science use cases across DOE science domains.
  - Provide the basis for organizing diverse program requirements into strategic priorities.
- 6 IRI Practice Areas represent critical topics that require close coordination to realize and sustain a thriving IRI ecosystem across DOE institutions.
  - Provide the basis for organizing the program governance model and cross-cutting efforts.

Convened over 150 DOE national laboratory experts from all 28 SC user facilities across 13 national laboratories to consider the technological, policy, and sociological challenges to implementing IRI.

Slide adapted from Ben Brown, ASCR

The 2022 IRI Architectural Blueprint Activity identified 6 key challenge areas and requirements from science teams

The IRI Framework comprises:

> 3 IRI Science Patterns represent integrated science use cases across DOE science domains.

Time-Sensitive Patterns

Data-Integration Patterns | Long Campaign Patterns

> 6 IRI Practice Areas represent critical topics that require close coordination to realize and sustain a thriving IRI ecosystem across DOE institutions.

Workflows, Interfaces & Automation | Scientific Data Lifecycle | User Experience

Portable/Scalable Solutions | Cybersecurity & Federated Access | Resource Co-Operations

Convened over 150 DOE national laboratory experts from all 28 SC user facilities across 13 national laboratories to consider the technological, policy, and sociological challenges to implementing IRI.

Slide adapted from Ben Brown, ASCR

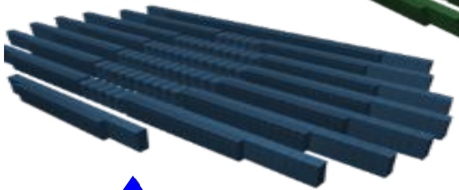
DOE has established an FY24-25 Agency Priority Goal to stand up the IRI program

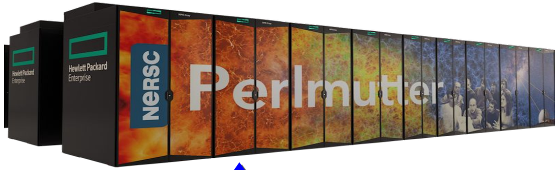
ASCR is implementing IRI through these major elements

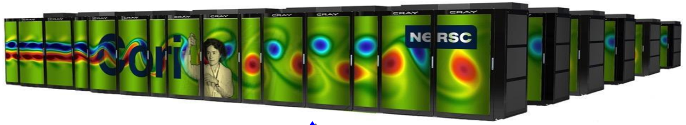
1. Invest in IRI foundational infrastructure
2. Stand up the IRI Program governance and FY24 workstreams
3. Bring IRI projects into formal coordination
4. Deploy an IRI Pathfinding Testbed across the four ASCR Facilities

Slide adapted from Ben Brown, ASCR

NERSC Systems Roadmap







NERSC-7: Edison
Multicore CPU

2013

NERSC-8: Cori
Manycore CPU
NESAP Launched: transition applications to advanced architectures

2016

NERSC-9: Perlmutter
CPU and GPU nodes
NESAP Expanded Simulation, Learning & Data: Continued transition of applications and support for complex workflows

2020

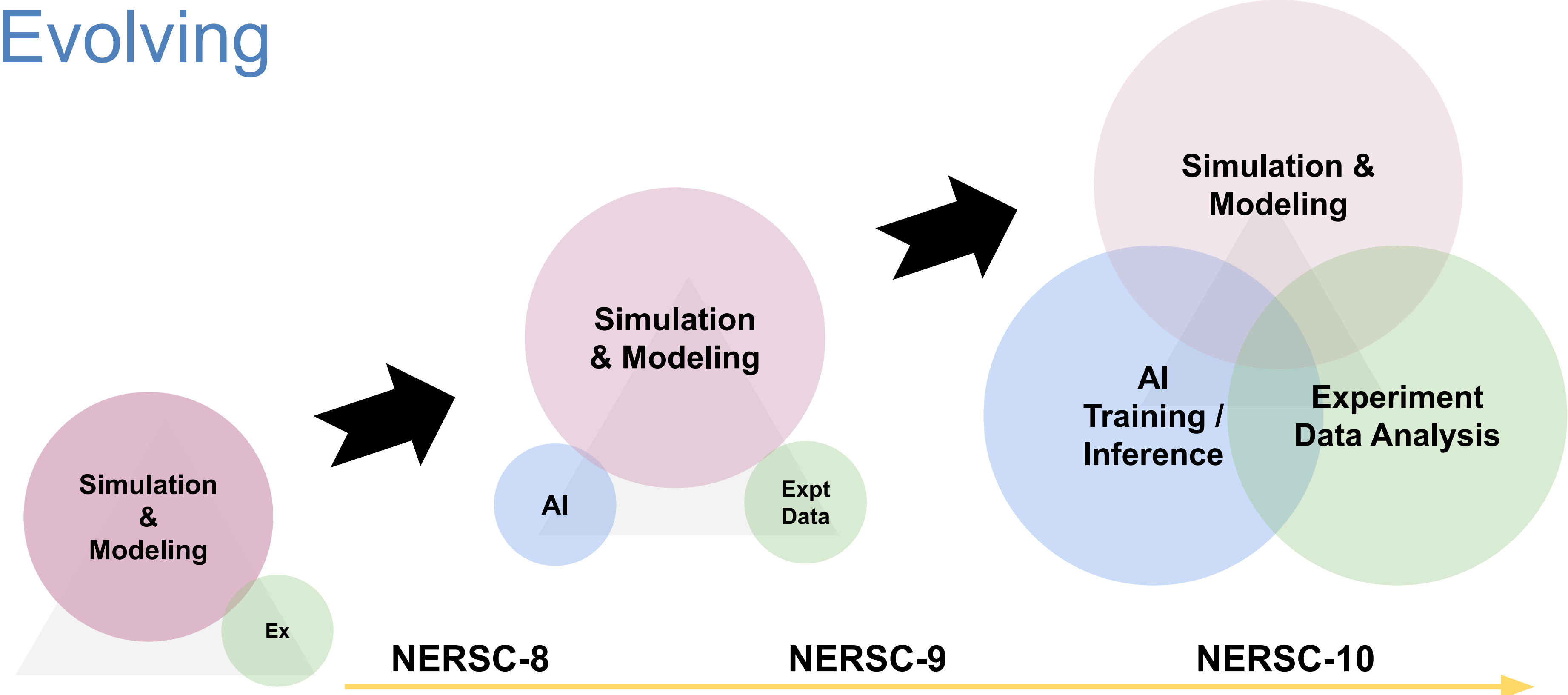
NERSC-10:
Exa system
NESAP Workflows: Accelerating end-to-end workflows with technology integration

2026

NERSC-11:
Beyond Moore

2030+

HPC Facility Workload Balance is Evolving



N10 User Requirements

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.**

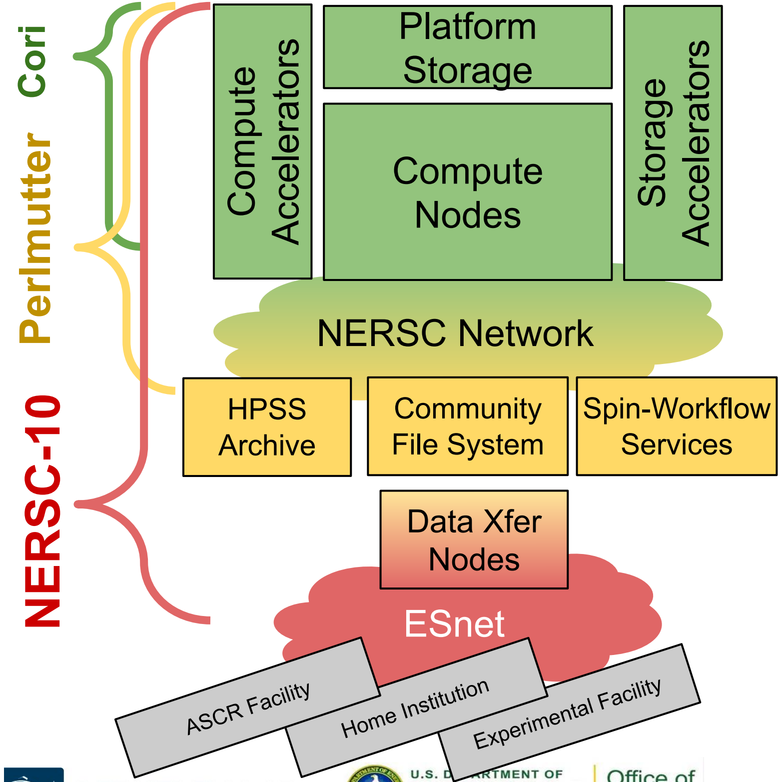
> ***NERSC-10 Mission Need Statement:***
> *The NERSC-10 system will **accelerate end-to-end** DOE SC **workflows** and enable new modes of scientific discovery through the integration of experiment, data analysis, and simulation.*

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

- **Quality of Service** – computation, storage and networking designed to emphasize response-time plus throughput/utilization.
- **Seamlessness** – tight integration of system components to enable high performance across workflow steps.
- **Portability** – Modular workflow execution across heterogeneous HPC, edge and cloud.
- **Programmability** – APIs to manage data, execute distributed code, and interact with system resources.
- **Orchestration** – coordinate resource management across different resource domains.
- **Security** – authentication, authorization and auditing (e.g., identify proofing, access/privacy control, records of transactions).

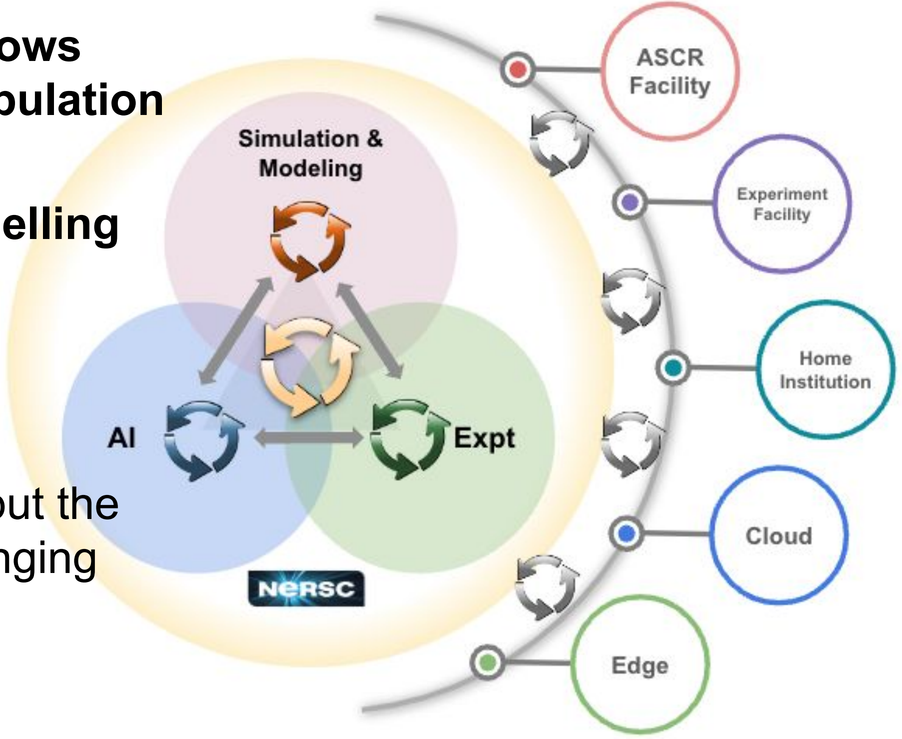


What is an HPC Workflow?

Workflows are interconnected computational and dataflow tasks with data products. They have task coupling (control flow) and/or data movement between tasks (data flow).

High performance computing (HPC) workflows interconnect computational and data manipulation steps across one/some/all of:

- **High performance simulation and modelling**
- **High performance AI workflows**
- **High performance data analytics**



We've been running workflows for decades - but the complexity and timeliness of workflows is changing which motivates a new approach with N10.

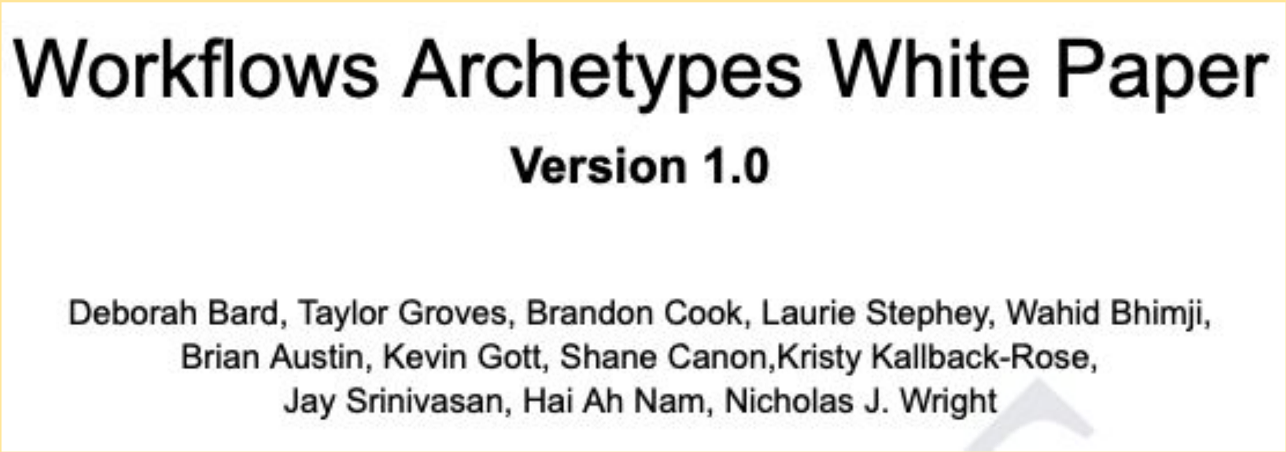
NERSC

We identified 6 workflows archetypes to help define our vision for N10

1. High-performance simulation & modeling workflow	large-scale multi-physics applications with checkpoint/restart, data post-processing, visualization
2. High-performance AI (HPAI) workflow	data integration-intensive science patterns such as training, inference, hyperparameter optimization
3. Cross-facility workflow: Rapid data analysis and real time steering	time-sensitive science patterns such as superfacility, edge, and hybrid cloud
4. Hybrid HPC-HPAI-HPDA workflow	long-term campaign science patterns, AI-in-the-loop, AI-around-the-loop
5. Scientific data lifecycle workflow: Interactive, data-analytics and viz	data integration-intensive science patterns such as Jupyter, scientific databases, VSCode
6. External event-triggered and API-driven workflow	time-sensitive science patterns such as function-as-a-service, microservices

We identified 6 workflows archetypes to help define our vision for N10

1. High-performance simulation & modeling wo	large-scale multi-physics applications with ... visualization
2. High-perfo	s such as ... ization
3. Cross-faci ... analysis and	uperfacility,
4. Hybrid HP	n-the-loop,
5. Scientific data ... Interactive, data-analytics and viz	erns such as Jupyter, scientific databases, VSCode
6. External event-triggered and API-driven workflow	time-sensitive science patterns such as function-as-a-service, microservices


Workflows Archetypes White Paper

Version 1.0

Deborah Bard, Taylor Groves, Brandon Cook, Laurie Stephey, Wahid Bhimji, Brian Austin, Kevin Gott, Shane Canon,Kristy Kallback-Rose, Jay Srinivasan, Hai Ah Nam, Nicholas J. Wright

search for “NERSC workflows white paper”

HPC Workflows Drive N10 Technology Capabilities

	Cloud native/ containers	QoS storage system (QSS)	End -to-end API	Network/ scheduling QoS	IRI/ Multi-site workflows	Smart networking	Prog. Env	Workflow Enablement Nodes (WEN, fka Spin)
1.Simulation & modeling		X	X			X	X	
2.AI	X	X	X	X	X	X	X	X
3.Cross-facility	X	X	X	X	X	X		X
4.Hybrid HPC-HPAI-HPDA	X	X	X	X	X	X	X	X
5.Scientific data lifecycle	X	X	X	X			X	X
6.Event-triggere d & API-driven	X	X	X	X		X	X	X

HPC Workflows Drive N10 Technology Capabilities

	Cloud native/ containers	QoS storage system (QSS)	End -to-end API	Network/ scheduling QoS	IRI/ Multi-site workflows	Smart networking	Prog. Env	Workflow Enablement Nodes (WEN, fka Spin)
1.Simulation & modeling		X	X			X	X	
2.AI	X	X	X	X	X	X	X	X
3.Cross-facility	X	X	X	X	X	X		X
4.Hybrid HPC-HPAI-HPDA	X	X	X	X	X	X	X	X
5.Scientific data lifecycle	X	X	X	X			X	X
6.Event-triggered & API-driven	X	X	X	X		X	X	X

Pink: cannot be done today
Orange: can be done only with extraordinary effort
Green: can be done today in limited way

Innovation in software is key to enabling complex workflows

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

Support usage of both ssh and Jupyter

Meet federal security requirements

User software/ workflows

Workflow environment

System software

System hardware

RESTful user-facing APIs support automation

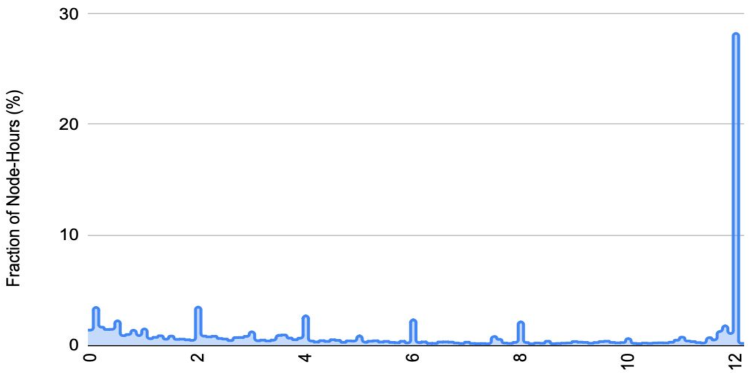
System-side APIs for workflow observability, administration and reconfigurability

Containerize the user environment

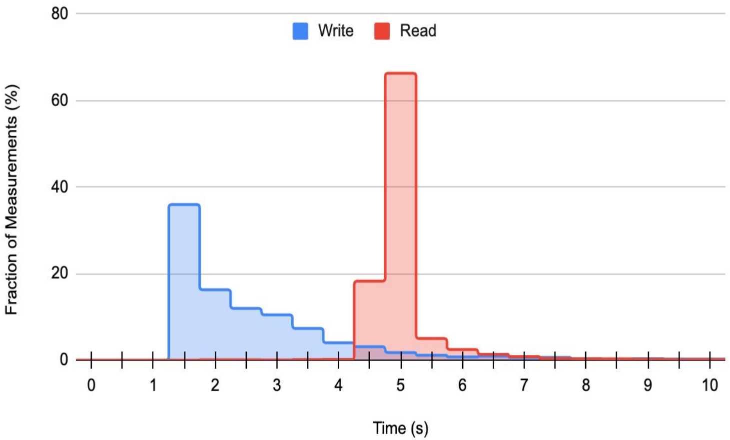
The NERSC workload requires capabilities that are hard to reconcile in a single file system

- 28% of all node hours are used by jobs that run to the wallclock limit (12 hours)
- Checkpointed applications can be preempted to support urgent compute needs

Job duration on Perlmutter



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

PSS and QSS will meet the needs of the full NERSC workload

- Platform Storage System (PSS) will meet the needs of much of the NERSC workload
  - Traditional parallel filesystem
  - Optimal for streaming I/O, checkpoint/restart
- Quality of Service Storage System (QSS) will provide controllable, guaranteed IOPs / bandwidth to meet the needs of time-sensitive workflows
  - Optimal for high IOPs workloads
  - Isolation from other workloads to eliminate perturbations

NERSC-10 RFP: Technical Requirements

search "nersc rfp"


September 15, 2023

RFP

Technical Requirements Document

for

NERSC-10 System

Version 3.0

DRAFT

Lawrence Berkeley National Laboratory is operated by the University of California for the U.S. Department of Energy under contract NO. DE-AC02-05CH11231.

1 RFP Technical Requirements Document for NERSC-10 System, Version 3.0, September 15, 2023

Technical Summary:

- No peak flops requirement
  - 10x on workflow component benchmarks
- CPU + GPU nodes
- Two kinds of storage
  - PSS - 120 PB, 20 TB/s
  - QSS - 80 PB, performance guarantees
- Workflow Environment (beyond the programming environment)
- Modular system software and management to support complex workflows

NERSC-10 Timeline

- Project Authorized by DOE (CD-0) - Sept 2021
- Advanced Acquisition Plan approved by DOE - March 2023
- **Draft RFP Release - 20 April 2023**

- Technical Design Review - August 2023
- Berkeley Lab Director's Review (Red Team) - Fall 2023
- **CD-1 - December 2023**

- **RFP Release - ~Feb 2024**
- **Contract signed (CD-2) - Late CY 2024**
- (Potential) Phase I or Pilot System- mid 2025
- Technical Decision Point - Late 2025
- **Main System Delivery - Late 2026**
- **User access 2027**

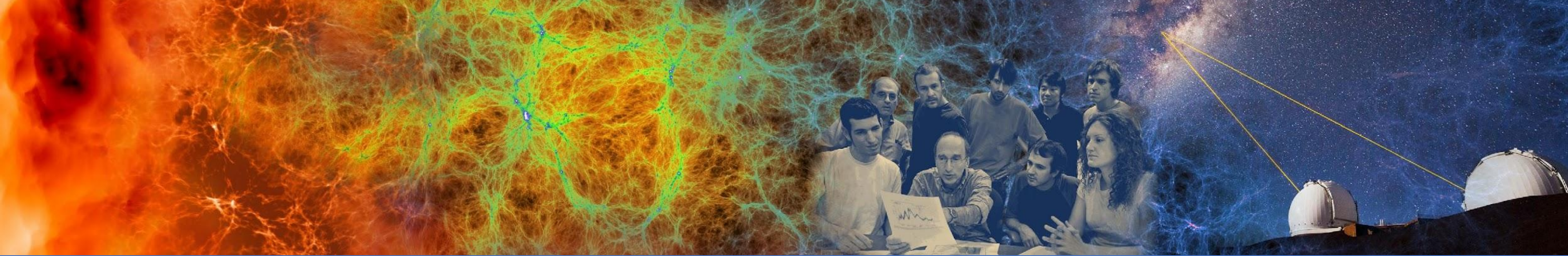
we are here

The NERSC-10 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 NERSC-10 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.

We're building an engagement model to coordinate a complex set of requirements and stakeholders in a changing technology landscape.

- *N10 will deliver 10x Perlmutter performance on HPC workflows.*
- *N10 is designed to be IRI-ready.*
- *GPU-enabled applications should have minimal issues in porting/running their applications.*
- *The N10 RFP will be released shortly, with system delivery in 2026.*



Thanks!

1 Invest in IRI foundational infrastructure

- IRI explicitly features in technical requirements/KPPs for next systems from NERSC & OLCF (to be deployed in ~2026).
- ESnet6 capabilities are already IRI-compatible.
- HPDF will be fully compatible with IRI

4 Deploy an IRI Pathfinding Testbed across the four ASCR Facilities

- Sometimes IRI will want to use a sandbox to develop services/ technologies necessary for IRI, eg for potentially disruptive services
- Every ASCR facility has designated hardware for this, deploy in FY24:
  - OLCF Advanced Computing Environment; NERSC Perlmutter On Demand; ALCF Edith; ESnet isolated wavelengths.

https://www.osti.gov/biblio/2205149
search for "Federated IRI Science Testbed (FIRST) Concept Note"

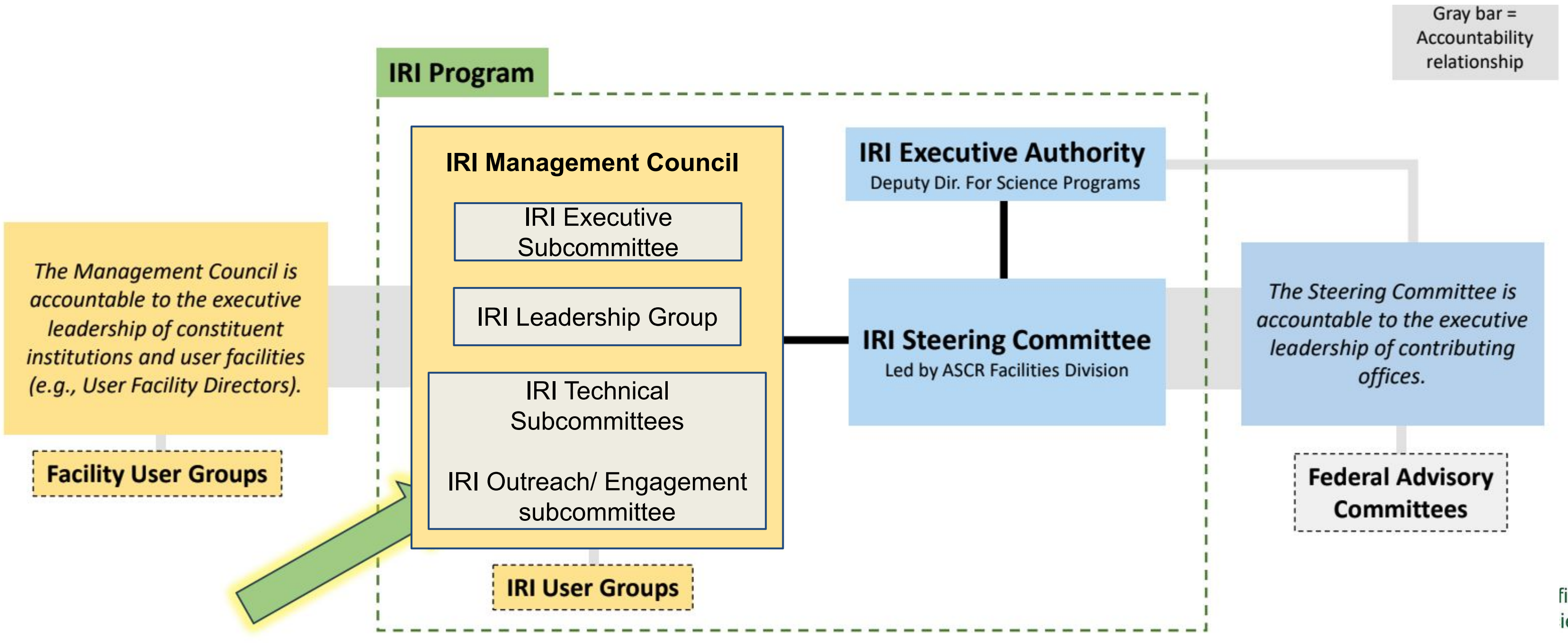
3 Bring IRI projects into formal coordination

- While governance model being figured out, IRI Working Group has been defining initial work plans
- Identified “pathfinder” science teams who are ready to move forward in partnership with IRI
  - *Requirements*; Identify any gaps in the technology areas proposed in the ABA report
  - Pathfinders nominated by DOE SC: **Lightsources, DIII-D, ESGF**
  - Other pathfinders: **LHC, JGI/KBase, GRETA/DELERIA, NCEM, …**
- Identified *initial* near-term actionable work areas (“workstreams”, “ABA practice areas”, …). Drafting charters for technical subcommittees:

- **Interfaces**
  - Facility API
  - Jupyter
- **Software Deployment and Portability**
  - containerization across sites
- **Security: authentication & access controls**
  - Federated ID
  - Provisioning Robot Accounts
- **Scheduling/preemption**
- **Data movement**
  - Globus, …

2 Stand up the IRI Program governance and FY24 workstreams

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



2 Stand up the IRI Program governance and FY24 workstreams

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

Active work happening now:

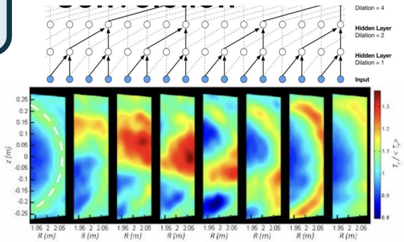
- Defining roles and responsibilities
- Defining initial work groups and plans
- Identifying ways to engage with the wider DOE community
- Writing IRI prospectus

Still very early days, but actively working on strategy for engagement outside of ASCR: there will be many opportunities to participate.

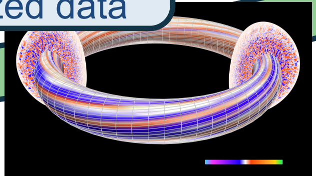
Example of Cross-facility Workflow: Fusion Experiment

Data readout, sent to NERSC

AI-driven data analysis



Simulation based on analyzed data



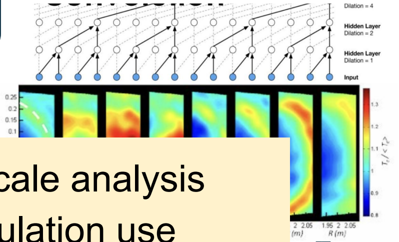
Feedback to scientist in minutes

Example of Cross-facility Workflow: Fusion Experiment

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

Data read
sent to NERSC

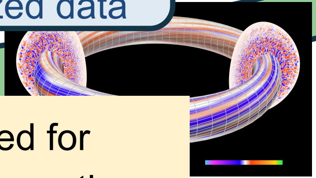
AI-driven data analysis



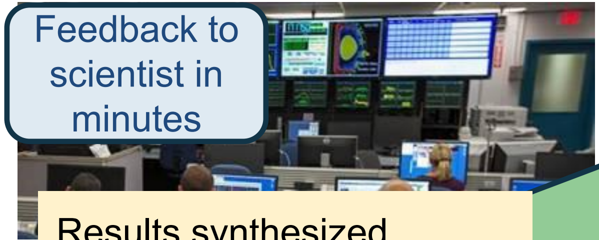
Data movement and compute progress tracked using **APIs** by automated workflow orchestrator and databases on **WENs**

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

Simulation based on analyzed data



Portable workflows designed for resiliency, possibly running on other resources if NERSC is unavailable



Feedback to scientist in minutes

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

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.

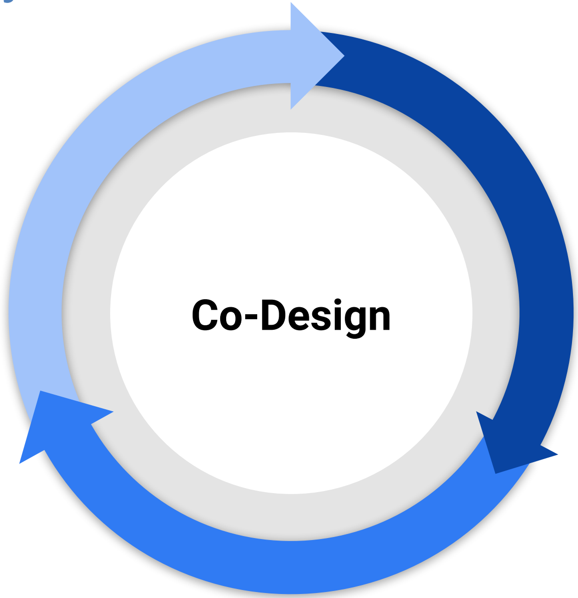
The N10 RFP is expected next year, system delivery in 2026

N10 will deliver 10x Perlmutter performance on HPC workflows

We will need close integration between our science teams, technology vendors, software providers and NERSC staff to ensure we meet the needs of the complex workflows of the future.

TECHNOLOGY INTEGRATION

Evaluation, development and integration of advanced tech; NRE



WORKFLOW READINESS

Code teams, IRI projects, vendors, and library/tools developers prepare for N10 workflows

WORKFLOW IMPLEMENTATION

Enabling high impact workflow capabilities & performance

NERSC

BERKELEY LAB
Bringing Science Solutions to the World

U.S. DEPARTMENT OF ENERGY | Office of Science