NERSc-10 and the Integrated Research Infrastructure (IRI)

HEP-CCE meeting, July 22nd 2024

Debbie Bard
Department Head,
Science Engagement and Workflows
NERSC

ASCAC Facilities charge report summary: Overarching recommendations (direct quotations)

Recommendation 1: Ensure the continued support and development of all five ASCR computational facilities reviewed—ALCF, OLCF, NERSC, HPDF, and ESnet—as they are central and essential to all SC science programs and broader national science and engineering research programs.

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 (henceforth referred to as 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.**

Recommendation 3: A comprehensive, coordinated R&D program delivering multiple prototype computing systems over a five-year timescale must be mounted to inform pathways for this integrated ecosystem, operational by 2034, due to (a) rapidly evolving economic and technical landscapes of the semiconductor and computing industries and (b) changing research practices.

"Recommendation 1 is necessary but not sufficient for success."

Slide adapted from Ben Brown, DOE

ASCAC Facilities charge report summary: IRI

"… the ASCR Ecosystem stands on its own to support national and international scientific, industry, and technological leadership, but there is an additional and very specific way that the ASCR ecosystem and its further integration is required specifically for SC.

"The Ecosystem serves, and will be essential to, the array of DOE SC experimental facilities that are generating exponentially growing data volumes, in what is referred to as the Integrated Research Infrastructure (IRI) …"

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

IRI is a new **program** encompassing all ASCR Facilities and participating Domain partners.
The IRI program involves the whole ASCR Facilities ecosystem (and potentially, the broader research computing ecosystem).

IRI is being designed to create better interfaces between DOE user facilities. It will include **frameworks for integration in software, policy and process**.

It will layer on top of existing (and future) facility capabilities, to enhance integration and usability/user experience.

Individual facilities will retain their missions and the ability to innovate in order to provide services to their user communities. IRI is the means by which we will manage "**the ASCR ecosystem and its further integration [that] is required specifically for SC**"

IRI Program launch is a DOE FY24-25 Agency Priority Goal
ASCR is implementing IRI through these four major elements

1 Invest in IRI foundational infrastructure

ASCR Facilities are architected to support IRI

- KPPs, benchmarks and readiness programs are explicitly designed for IRI

2 Stand up the IRI Program governance and technical activities

The IRI Management council has been established, and is working to:

- Define roles, responsibilities and reporting structures
- Launch committees to coordinate technical and outreach work

3 Bring IRI projects into formal coordination

We are coordinating existing work and forging new collaborations

- Roadmap for partnering with science teams
- Coordinating with CS research programs

4 Deploy an IRI Science Testbed across the ASCR Facilities

Shared development environment for IRI developers and pilot applications, designed to accelerate transition to operations

Standup of the IRI Program is a DOE FY24-25 Agency Priority Goal

FY 2021 President's Budget Request includes **Integrated** Computation and Data Infrastructure Initiative

ASCR IRI Task Force launch

ASCR IRI Task Force report

SC IRI Blueprint Activity launch

IRI Blueprint Activity results

FY 2024 PBR advances IRI and the High Performance Data Facility

HPDF Selection

GO

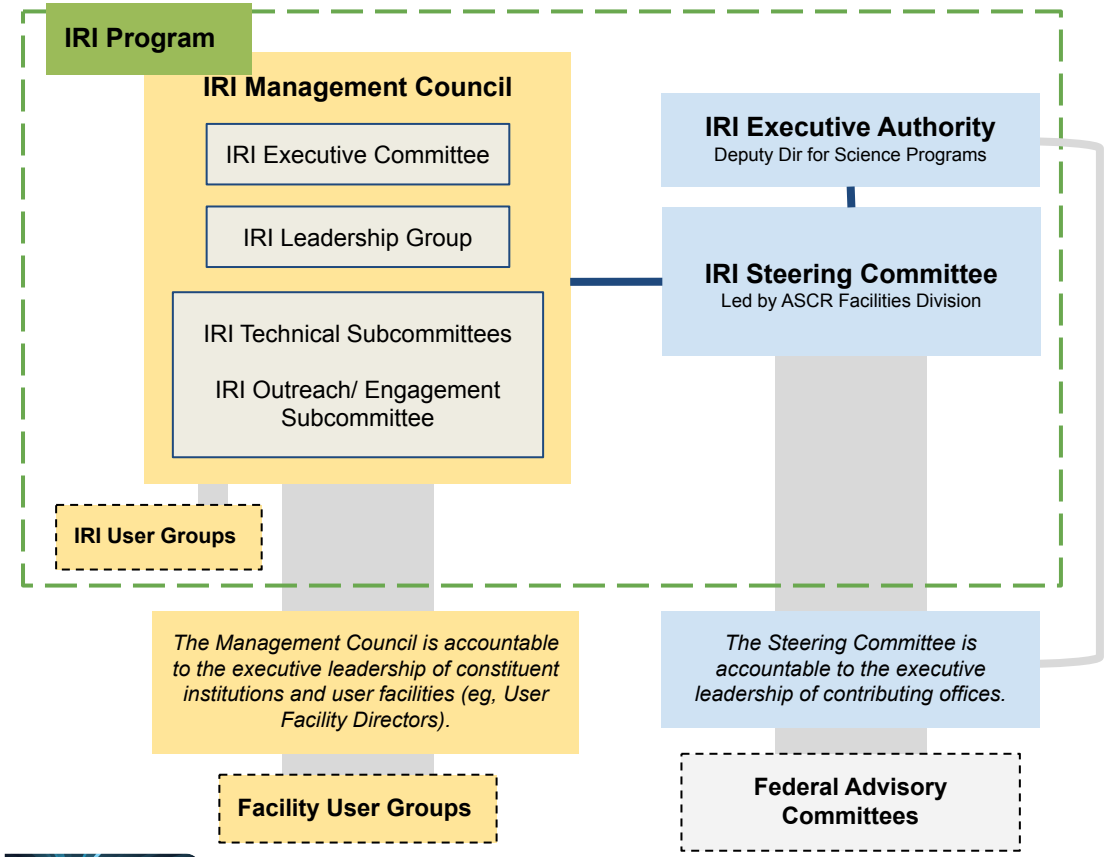
IRI Program Development

Jan **2020** Jan **2021** Jan **2022** Jan **2023** Jan **2024**

Vision **Strategy** **Implement**

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

The IRI Governance model



This structure was developed by the SC Headquarters IRI Coordination Group in 2022-23.

In late 2023, ASCR directed the ASCR Facilities to jumpstart the IRI Management Council.

The IRI program is moving out of the initial stages of development and definition

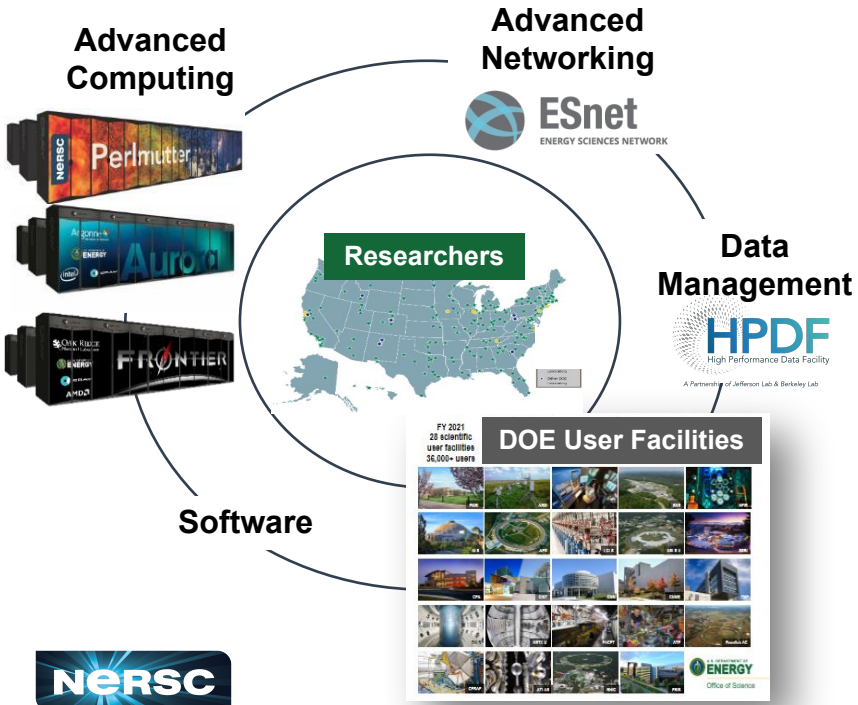
We are launching first phases of science engagements and technical work

- We've built the initial organizational structure to manage IRI efforts and integrate across program offices
- Now we're ready to develop our partnerships and lay out the pathways to engage with SC

We are still at the very early stages of the program. Last week, we held the joint IRI/HPDF meeting to kick off the expansion of IRI to broader partnerships. The outcome of that meeting will define the IRI roadmap, and guide how domain partners are brought into the program.

The ASCR Facilities are firmly embedded in the IRI ecosystem

IRI is permeating everything we do - **multi-facility workflows** are an integral part of our major infrastructure design and strategic planning.

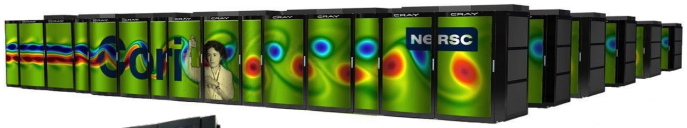


- IRI requires an exponentially growing amount of collaboration across ASCR facilities - it's changing how we work together
  - Discussed every week at the ASCR Facility Directors meeting
  - Catalysing cross-facility events on everything from training to API design
- ASCR facilities have made a lot of progress in supporting complex workflows
- Ongoing projects, testbeds, and major acquisitions are targeting IRI (ESnet6, NERSC-10, OLCF-6, HPDF and ALCF-4)

NERSC Systems are becoming more energy efficient

NERSC-7: Edison
Multicore CPU
- **2.6 PFlops**
- **~2.5MW**

2013

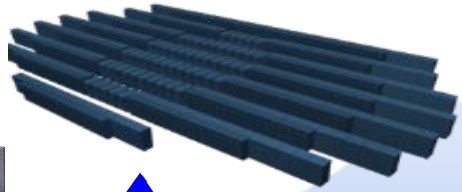


NERSC-8: Cori
Manycore CPU
- **27 PFlops**
- **~4 MW**

2016

NERSC-9: Perlmutter
CPU and GPU nodes
- **113 PFlops**
- **~6 MW**

2020

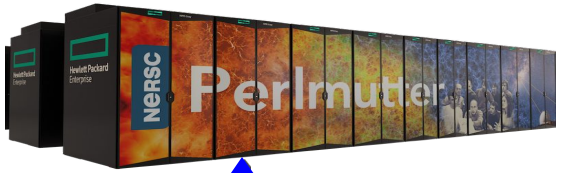


NERSC-10:
Exa system
- **Exaflop???**
- **<20MW**

2026

Largely a function of more energy-efficient chips and cabinet cooling improvements

NERSC is responding to the evolving DOE Mission

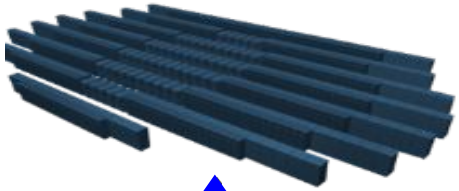


NERSC-8: Cori
Manycore CPU architectures

2016

NERSC-9: Perlmutter
CPU and GPU nodes
Expanded Simulation, Learning & Data

2020

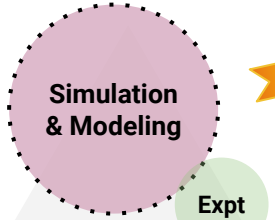


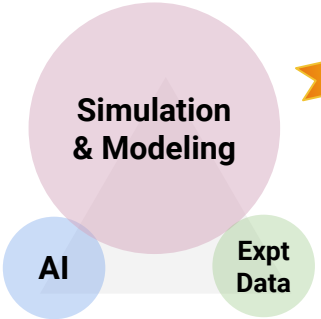
NERSC-10:
Accelerating end-to-end workflows

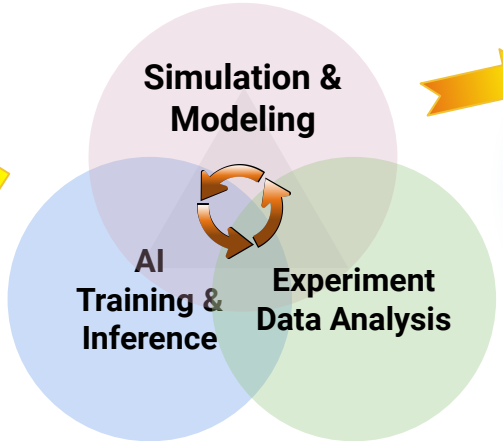
2026

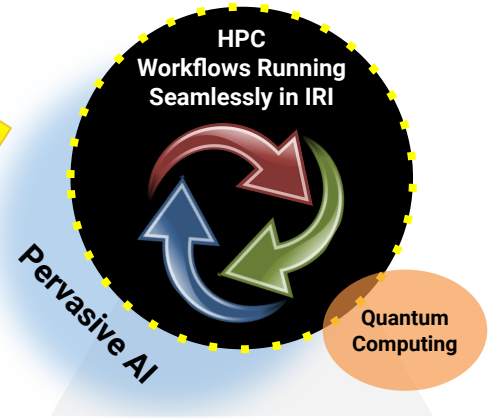
NERSC-11:
Beyond Moore

2030+





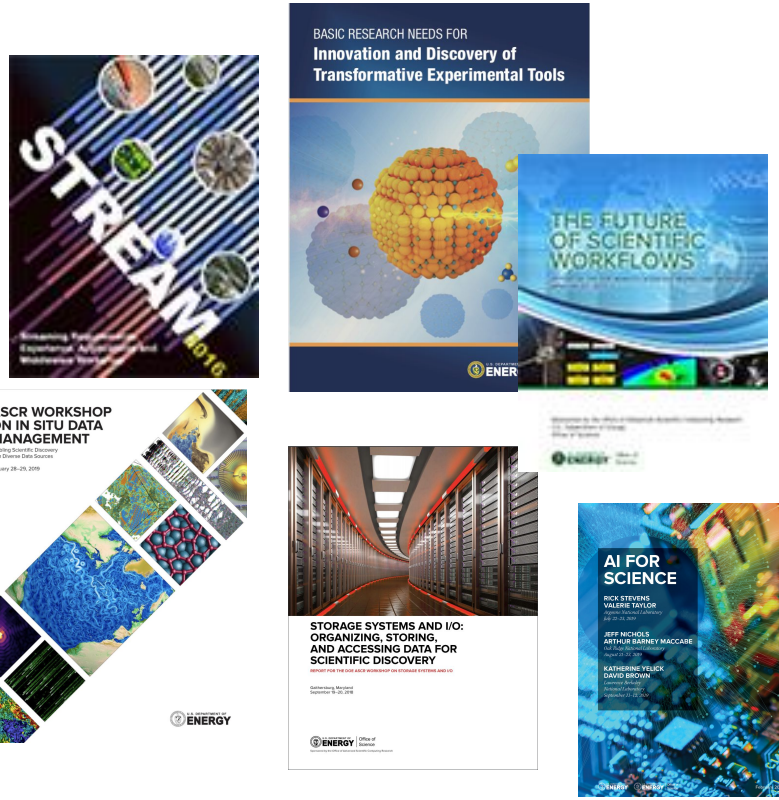




NERSC-10 requirements reflect the DOE Mission

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

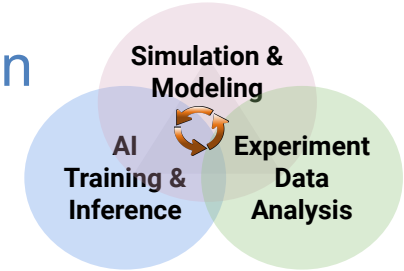
Users need the ability to search, analyze, reuse, and combine data from different sources into **large scale simulations and AI models.**



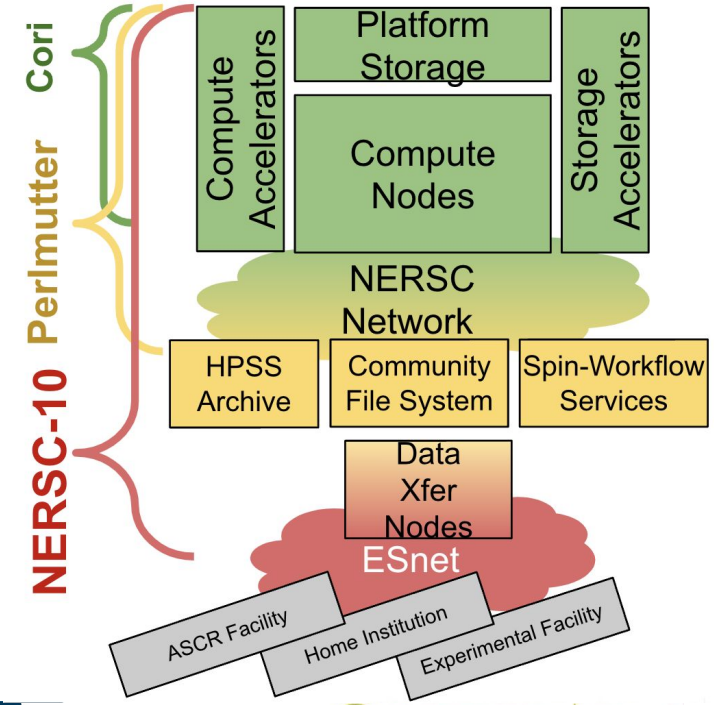
NERSC-10 Mission Need Statement:

*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.*

NERSC-10 is Designed to Support Complex Simulation and Data Analysis Workflows at High Performance



- **Quality of Service**: computation, storage and networking enables response-time plus utilization.
- **Seamlessness**: tight integration of system. components enables high performance workflows.
- **Programmability**: APIs manage data, execute code, and interact with system resources.
- **Orchestration**: coordinates resource management across domains.
- **Portability:** Modular workflow execution across IRI sites.
- **Security:** authentication, authorization and auditing.

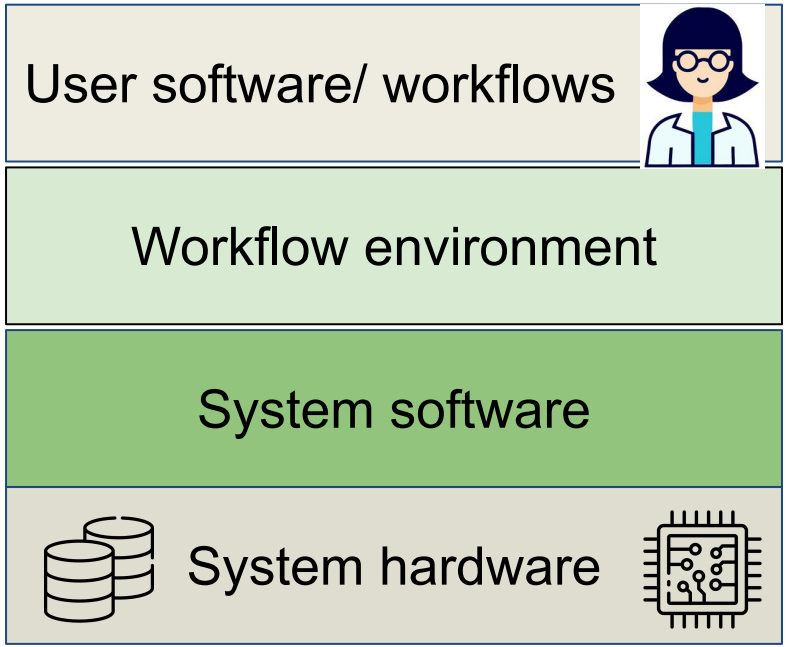


Innovation in software is key to enabling complex IRI workflows

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

Support usage of both ssh and Jupyter

Meet federal security requirements



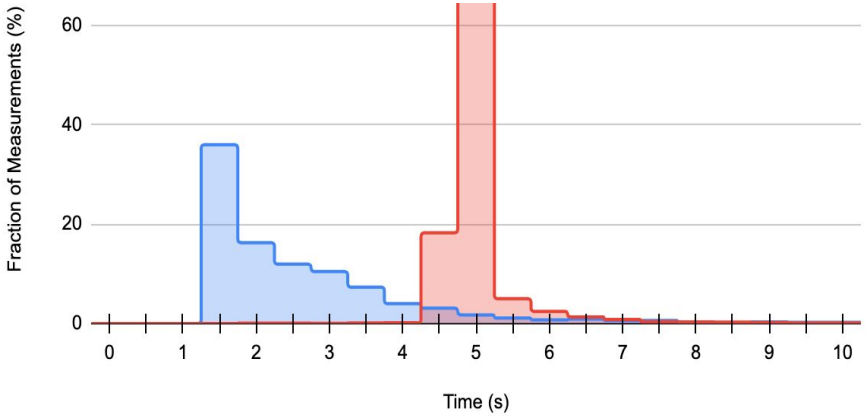
RESTful user-facing APIs support automation

System-side APIs for workflow observability, administration and reconfigurability

Containerize the user environment

The NERSC workload is hard to reconcile with a single file system

IOR performance on Perlmutter



- 2% of all IOR write tests on Perlmutter took at >5x longer than the mode

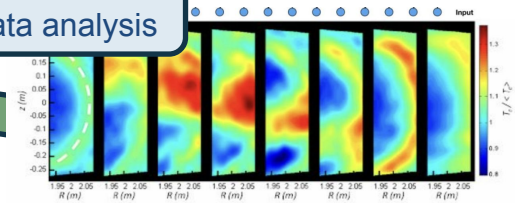
For instrument-driven and time-dependent workflows such variance is seriously disruptive.

- N10 will have a **Quality of Service Storage System (QSS)** to provide controllable, guaranteed IOPs / bandwidth for time-sensitive workflows
- Platform Storage System (PSS) is a more traditional FS that will meet the needs of much of the NERSC workload

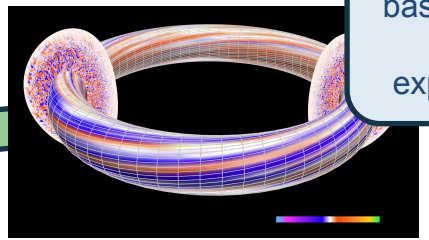
Cross-facility workflow example: Fusion science with DIII-D, preparing for ITER

Data readout from tokamak sensors sent to NERSC

AI-driven data analysis

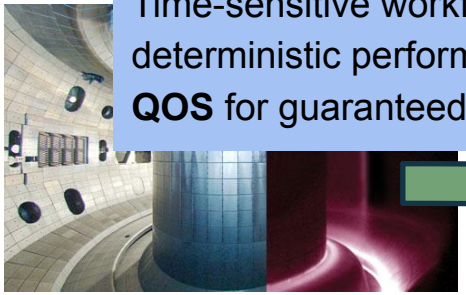


Digital Twin simulation based on analyzed data: recommends new experiment parameters



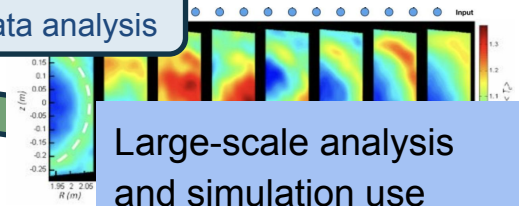
Feedback to scientist in minutes

Cross-facility workflow example: Fusion science with DIII-D, preparing for ITER



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

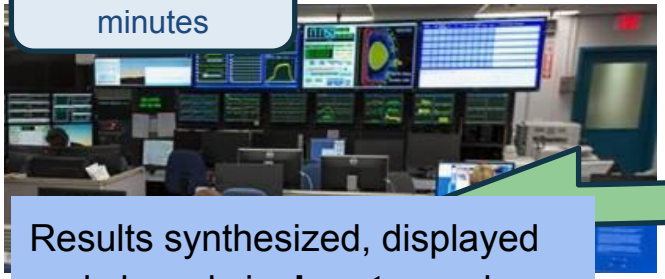
AI-driven data analysis



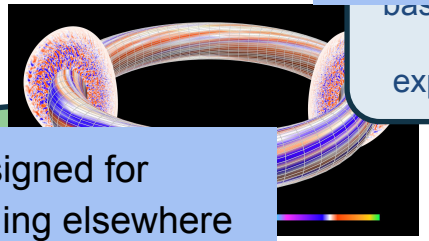
Data movement and compute progress tracked using **APIs** by automated workflow orchestrator

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

Feedback to scientist in minutes



based on analyzed data. recommends new experiment parameters



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

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

NESAP is Evolving for NERSC-10

NESAP for Perlmutter

"Application" performance **Tuning kernels, communication, maximize system performance**

- NESAP for Simulations (algorithms, data structures, GPU, parallelism, scalability)

- NESAP for Data (data intensive science pipelines)
- NESAP for Learning (machine learning and deep learning solutions)

Workflow Performance

Workflows may involve all of Simulation, Data and Learning activities, including I/O, system optimizations, data collection and transfer, training, etc.

Workflow Capabilities

Includes codesign, implementation and demonstration of Advanced Technologies Capabilities

Two NESAP Tracks: Agile Engagement

Workflow Pathfinding track
New workflow technologies often start in the user community => **Flexible program, where projects are refreshed annually**.

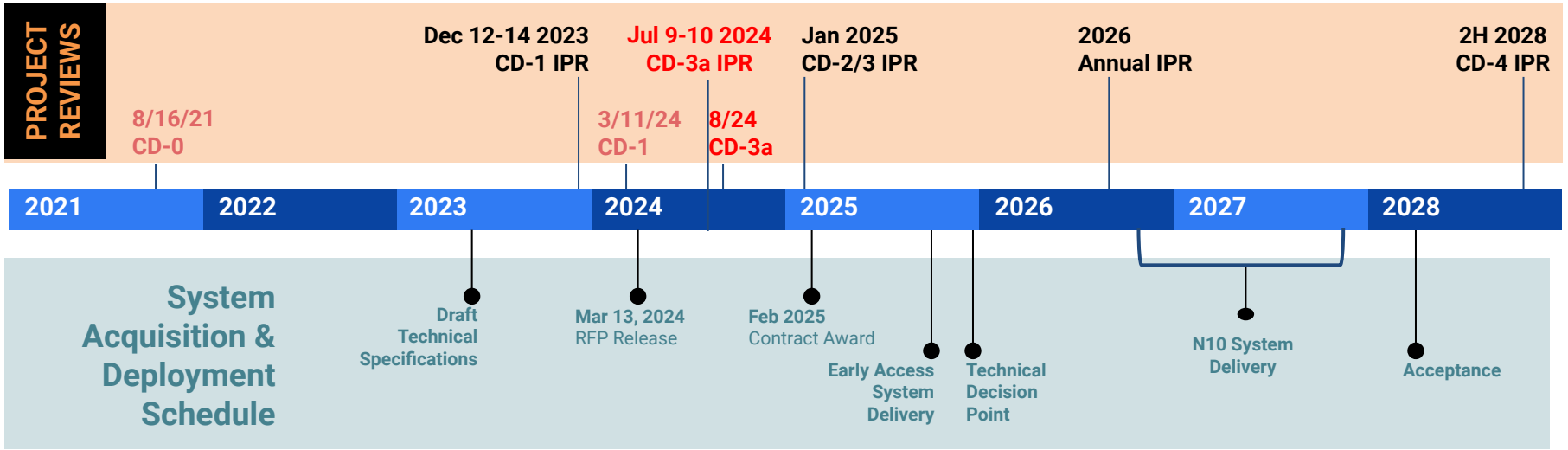
Focus on pathfinding for next-generation user workflows

Strategic Projects track
Addressing technology gaps requires long-term strategic engagements => **Multi-year program from CD-2 till system acceptance.**

Focused on N10 deployment, KPPs, & co-design

NERSC-10 Schedule

- We aim to provide user access to NERSC-10 in 2026/27 to maximize user productivity and ensure sufficient time to migrate our users before Perlmutter is decommissioned (EOL approx. 2027/2028).



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 our decades of work with the experiment science community and IRI.

- *N10 will deliver 10x Perlmutter performance on HPC workflows.*
- *N10 is designed to be IRI-ready.*