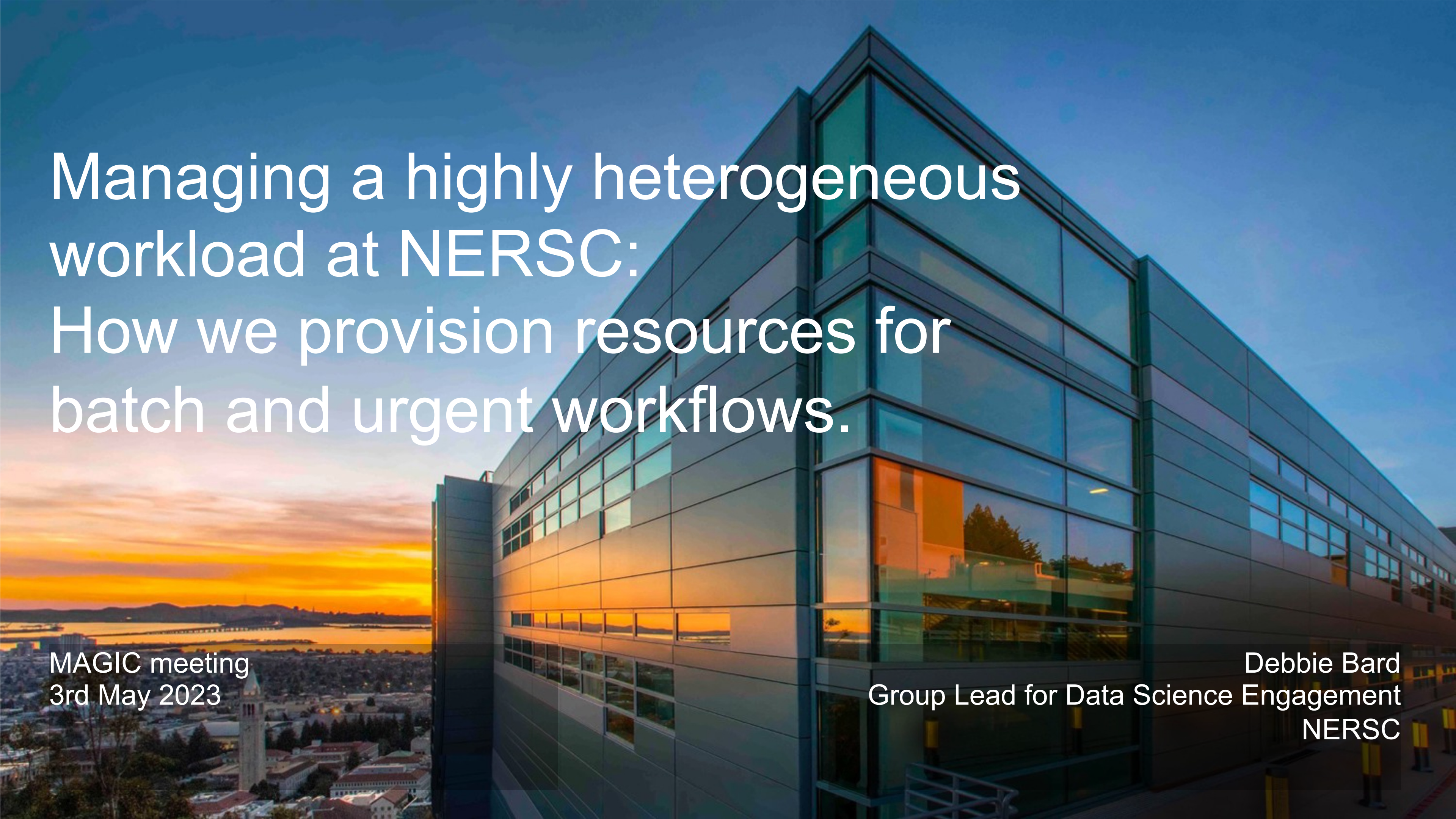


Managing a highly heterogeneous workload at NERSC: How we provision resources for batch and urgent workflows.

MAGIC meeting
3rd May 2023

Debbie Bard
Group Lead for Data Science Engagement
NERSC

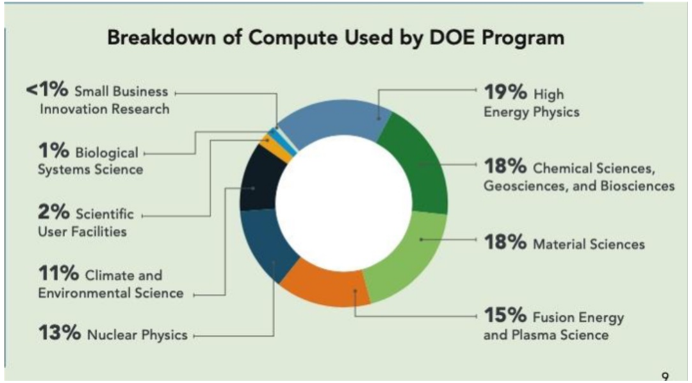
NERSC is the mission High Performance Computing facility for the DOE Office of Science

9,000 Users
1,000 Projects

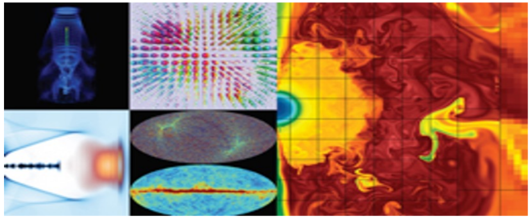
nature
JACS
Cell Reports
Physical Science
nature materials
Suppressing platinum dissolution
nature physics
Blinded by the light neutrino
JCTC Journal of Chemical Theory and Computation
ACS Publications

>2,000

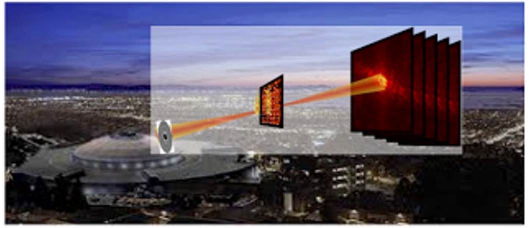
Scientific Journal Articles per Year



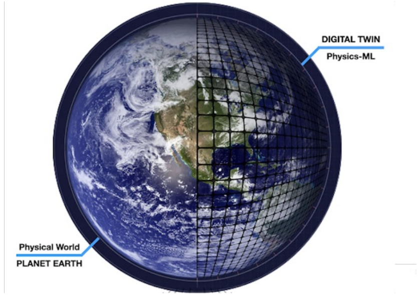
NERSC BY THE NUMBERS
2021 NERSC USERS ACROSS US AND WORLD
50 States + Washington D.C. and Puerto Rico
46 Countries



Simulations at scale

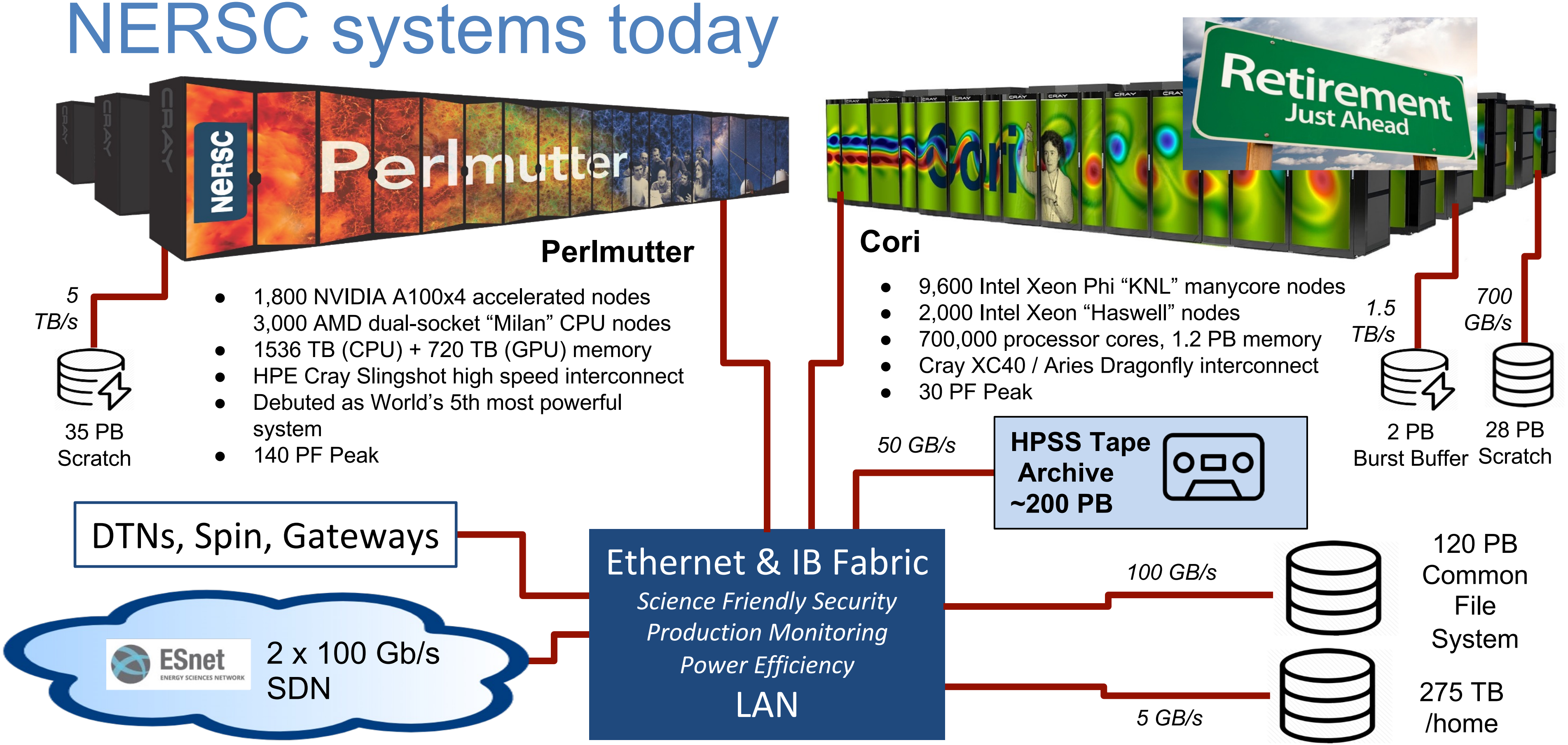


Urgent and interactive computing
Photo Credit: CAMERA



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

NERSC systems today



Perlmutter

- 1,800 NVIDIA A100x4 accelerated nodes
  3,000 AMD dual-socket "Milan" CPU nodes
- 1536 TB (CPU) + 720 TB (GPU) memory
- HPE Cray Slingshot high speed interconnect
- Debuted as World's 5th most powerful system
- 140 PF 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

HPSS Tape Archive ~200 PB

DTNs, Spin, Gateways

2 x 100 Gb/s SDN

Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
LAN

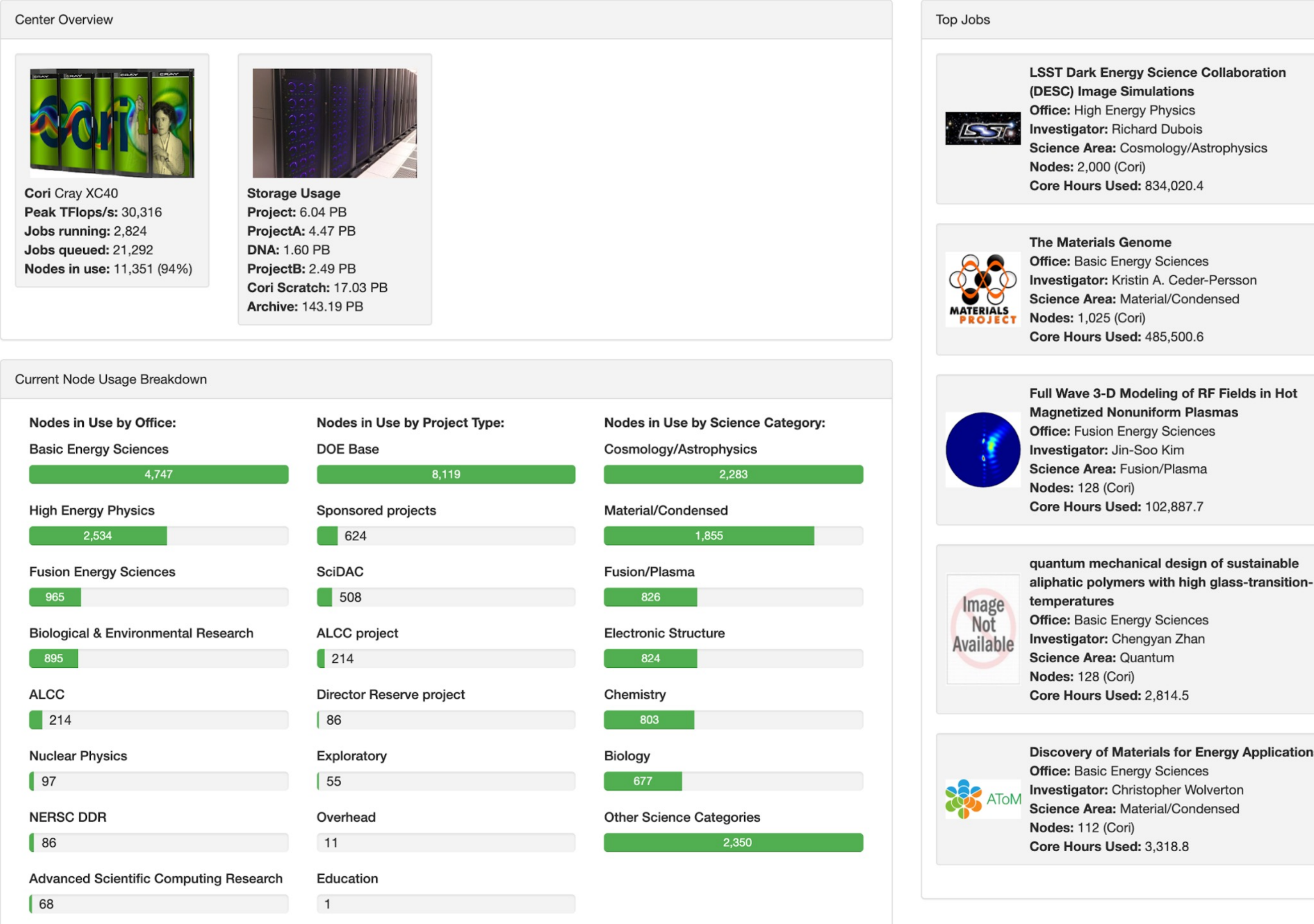
120 PB Common File System

275 TB /home

NERSC has a large and diverse workload

Snapshot of live computing:

- 2824 jobs running
- 21,292 jobs queued
- 94% utilization
- Mixture of simulation and data analysis



Center Overview

Cori Cray XC40
Peak TFlops/s: 30,316
Jobs running: 2,824
Jobs queued: 21,292
Nodes in use: 11,351 (94%)

Storage Usage
Project: 6.04 PB
ProjectA: 4.47 PB
DNA: 1.60 PB
ProjectB: 2.49 PB
Cori Scratch: 17.03 PB
Archive: 143.19 PB

Current Node Usage Breakdown

Nodes in Use by Office:		Nodes in Use by Project Type:		Nodes in Use by Science Category:	
Basic Energy Sciences	4,747	DOE Base	8,119	Cosmology/Astrophysics	2,283
High Energy Physics	2,534	Sponsored projects	624	Material/Condensed	1,855
Fusion Energy Sciences	965	SciDAC	508	Fusion/Plasma	826
Biological & Environmental Research	895	ALCC project	214	Electronic Structure	824
ALCC	214	Director Reserve project	86	Chemistry	803
Nuclear Physics	97	Exploratory	55	Biology	677
NERSC DDR	86	Overhead	11	Other Science Categories	2,350
Advanced Scientific Computing Research	68	Education	1		

Top Jobs

LSST Dark Energy Science Collaboration (DESC) Image Simulations
Office: High Energy Physics
Investigator: Richard Dubois
Science Area: Cosmology/Astrophysics
Nodes: 2,000 (Cori)
Core Hours Used: 834,020.4

The Materials Genome
Office: Basic Energy Sciences
Investigator: Kristin A. Ceder-Persson
Science Area: Material/Condensed
Nodes: 1,025 (Cori)
Core Hours Used: 485,500.6

Full Wave 3-D Modeling of RF Fields in Hot Magnetized Nonuniform Plasmas
Office: Fusion Energy Sciences
Investigator: Jin-Soo Kim
Science Area: Fusion/Plasma
Nodes: 128 (Cori)
Core Hours Used: 102,887.7

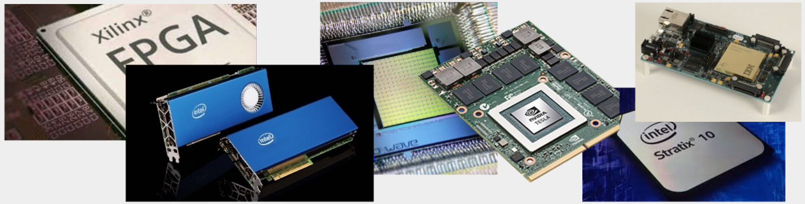
quantum mechanical design of sustainable aliphatic polymers with high glass-transition-temperatures
Office: Basic Energy Sciences
Investigator: Chengyan Zhan
Science Area: Quantum
Nodes: 128 (Cori)
Core Hours Used: 2,814.5

Discovery of Materials for Energy Application
Office: Basic Energy Sciences
Investigator: Christopher Wolverton
Science Area: Material/Condensed
Nodes: 112 (Cori)
Core Hours Used: 3,318.8

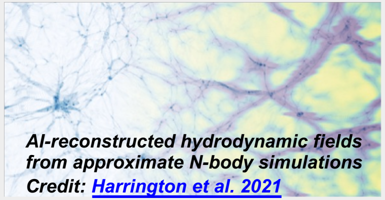
A changing computing landscape challenges us to think differently about supporting the Office of Science workload

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

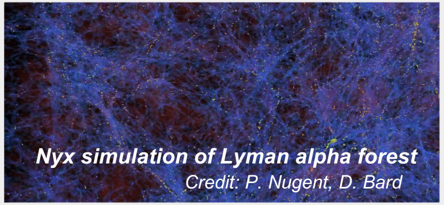
The proliferation of accelerators and **new technologies**



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



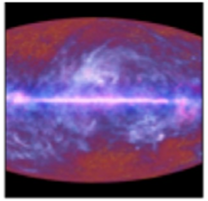
AI-reconstructed hydrodynamic fields from approximate N-body simulations
Credit: [Harrington et al. 2021]



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

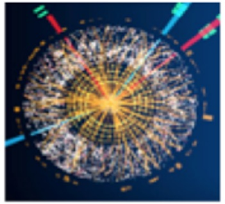
NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities

Palomar Transient Factory Supernova



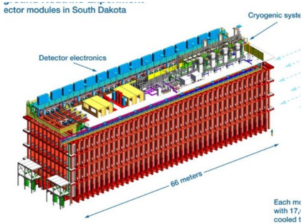
Planck Satellite Cosmic Microwave Background Radiation

Star Particle Physics

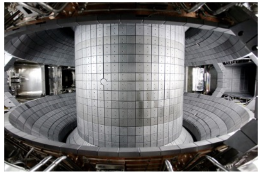


Atlas Large Hadron Collider

APS



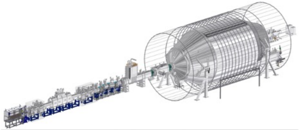
Dune



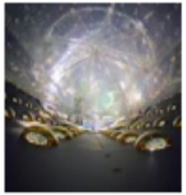
KStar

ARM

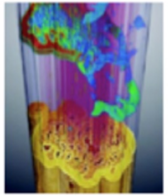
GlueX



Katrin



Dayabay Neutrinos

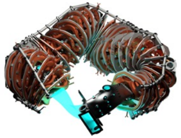


ALS Light Source

LCLS Light Source

Joint Genome Institute Bioinformatics

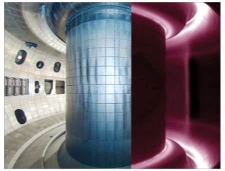
NSLS-II



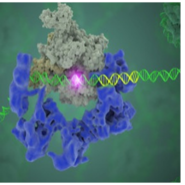
HSX

Majorana

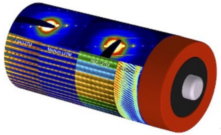
AmeriFlux



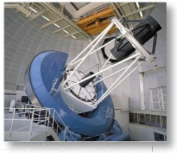
DIII-D



Cryo-EM

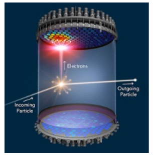


NCEM



DESI

LSST-DESC



LZ

IceCube

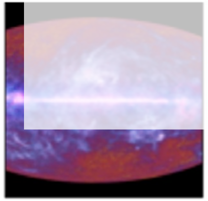
EXO

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

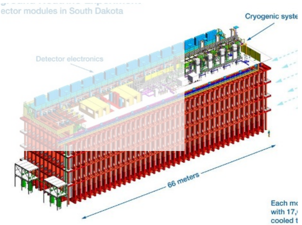


Planck Satellite Cosmic Microwave Background Radiation

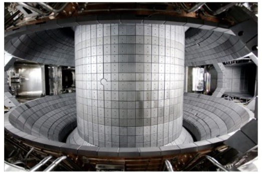
Star Particle Physics

Atlas Large Hadron Collider

APS



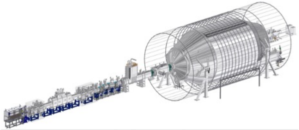
Dune



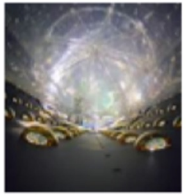
KStar

ARM

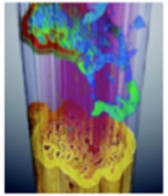
GLUEX



Katrin



Dayabay Neutrinos

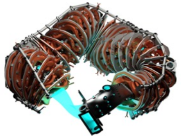


ALS Light Source

LCLS Light Source

Joint Genome Institute Bioinformatics

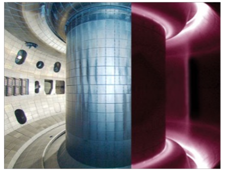
NSLS-II



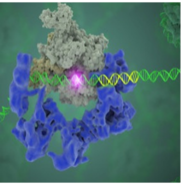
HSX

Majorana

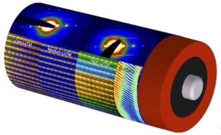
AMERIFLUX



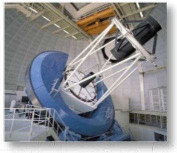
DIII-D



Cryo-EM

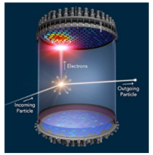


NCEM



DESI

LSST-DESC



LZ

IceCube

EXO

JBEI
Joint BioEnergy Institute

Requirements reviews and users from experimental facilities describe numerous pain points

- **Workflows** require manual intervention and custom implementations
- Difficult to surge experimental pipelines at HPC facility in ‘**real-time**’
- I/O performance, storage space and access methods for **large datasets** remain a challenge
- Searching, publishing and sharing **data** are difficult
- **Analysis codes** need to be adapted to advanced architectures
- Lack of **scalable analytics software**

Technical

- **Resilience strategy** needed for fast-turnaround analysis
  - including: coordinating maintenances, fault tolerant pipelines, rolling upgrades, alternative compute facilities...
- No **federated identity** between experimental facilities and NERSC
- Not all scientists want command-line access.

Policy

Requirements reviews and users from experimental facilities describe numerous pain points

- **Workflows** require manual intervention and custom implementations
- Difficult to surge experimental pipelines at HPC facility in ‘**real-time**’
- I/O performance, storage space and access methods for **large datasets** remain a challenge
- Searching, publishing and sharing **data** are difficult
- **Analysis codes** need to be adapted to advanced architectures
- Lack of **scalable analytics software**

Technical

- **Resilience strategy** needed for fast-turnaround analysis
  - including: coordinating maintenances, fault tolerant pipelines, rolling upgrades, alternative compute facilities...
- No **federated identity** between experimental facilities and NERSC
- Not all scientists want command-line access.

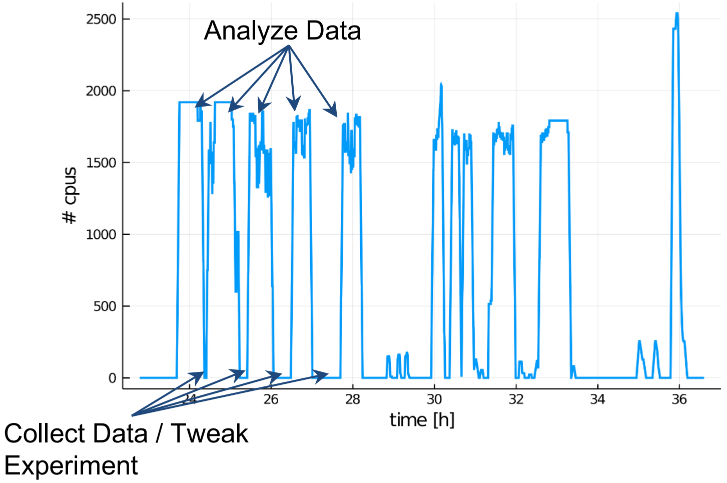
Policy

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 without damaging system utilization?

- 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



Large-scale simulations
“I need to run my climate model on 9000 nodes”

Shared-node Queue
Transfer Queue
“I only need 2 cores (and I’m not willing to pay for the full node)”
“I need to transfer many PB of data”

Pipeline/Workflow Management Nodes
“I need a service running 24/7 to manage my data analysis pipeline”

Real-Time Queues for Co-Scheduling w/Experiments
“I need to analyse this microscope data immediately otherwise my experiment will fail”

slurm
workload manager

Deadline Computing
“I need to analyze this telescope data before sunrise”

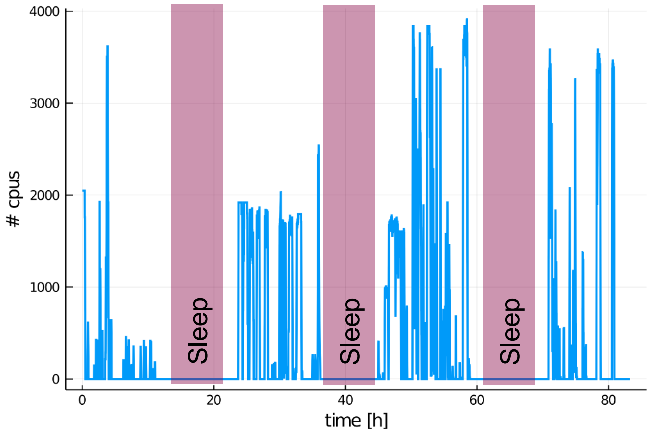
Interactive Queues:
4 Nodes x 4 Hours
“I need to interactively debug my code at scale without waiting in the batch queue”

Realtime queue

- User requests access to the “Realtime” qos via a form
  - Frequency, # nodes, job length, reason
- Small number of nodes “reserved” for fast-start jobs, and these jobs enter the queue with very high priority
  - Typically start within a few minutes – advantage of large number and mix of jobs on our systems
  - But we don’t let them disrupt the start time of a large job for which we have been draining the system, so we can maintain utilization
- This works well for small jobs (~tens of nodes) which covers many of our use cases

Reservations for urgent computing

Use reservations (can be hundreds of nodes) to guarantee compute will be available during shifts



Slide adapted from Johannes Blaschke

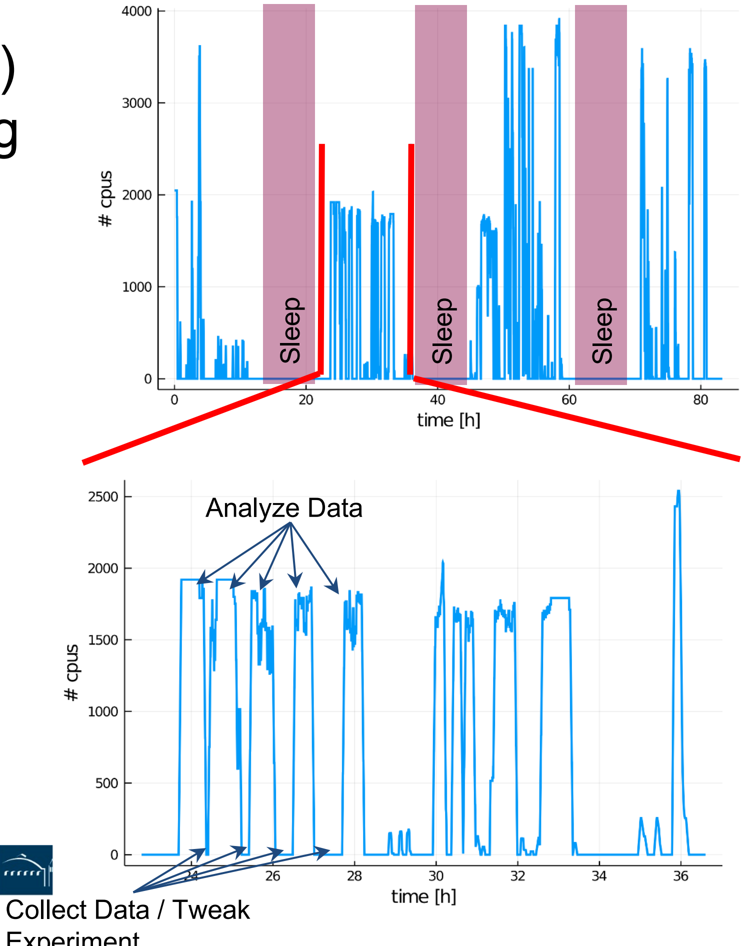
Reservations for urgent computing

Use reservations (can be hundreds of nodes) to guarantee compute will be available during shifts

But during a shift, sometimes the reserved compute nodes sit idle

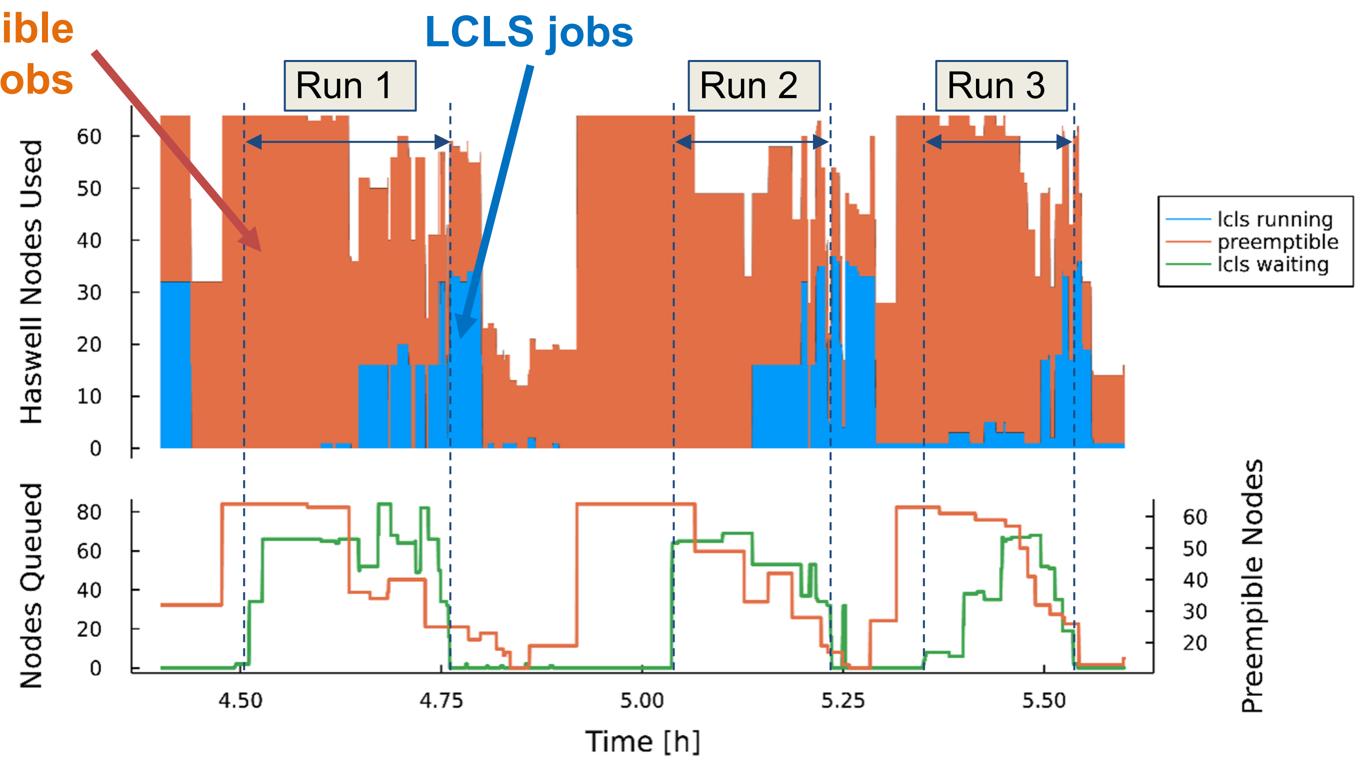
- Adjust sample
- Adjust experiment parameters
- Deal with problems
- Eat lunch

→***Bursty compute needs***



Slide adapted from Johannes Blaschke

Preemptible jobs fill the “gaps” in reservations



Preempt Queue: qos from Slurm

- Jobs in this queue can be preempted in the favor of a higher priority job.
- Jobs can be requeued.
- Your application must have checkpoint restart capabilities to take advantage of this.
- A typical use case would be a job that requires very long time to complete (it may take very long for it to schedule without preempt queue).

```
#SBATCH -q preempt
#SBATCH -C gpu
#SBATCH -N 1
#SBATCH --time=24:00:00
#SBATCH --error=%x-%j.err
#SBATCH --output=%x-%j.out
#SBATCH --comment=96:00:00  #desired time limit
#SBATCH --signal=B:USR1@60  #sig_time (60 seconds) checkpoint overhead
#SBATCH --requeue
#SBATCH --open-mode=append
```

We’re still working on figuring out how to get the charging right – subtracting charged preemptible jobs from the reservation post-hoc requires new tooling in our account management system.

Developing Transparent Checkpointing

- NERSC is engaged in development work with Northeastern University researchers and MemVerge, Inc. to improve, test, and deploy transparent user-space checkpoint-restart tools
- Distributed MultiThreaded CheckPointing (DMTCP) and it’s plugins such as MPI-Agnostic, Network-Agnostic MPI (MANA) can conveniently add checkpointing wrappers to workloads that don’t otherwise include it
- We’ve focused on VASP as the model application, which by itself makes at least 20% of the NERSC workload suitable for the the Preempt QOS
  - Also looking at incorporating DMTCP checkpointing into workflow orchestrators like gnuparallel…

Credit: Bill Arndt (NERSC)

Experiment	Science case	Time frame	Urgency	Job scale	Method
Linac Coherent Light Source LCLS	Rapid data analysis to guide running experiment	12-hour shift scheduled months in advance, bursty use of NERSC during shift	Minutes	100s-1000s nodes	Real-time and Reservations
Dark Matter detection LZ	Continuous monitoring of detector health	24/7	Minutes/ hours	<10 nodes (100 during calibration runs)	Realtime (reservation for calibration)
National Center for Electron Microscopy MOLECULAR FOUNDRY	Rapid data analysis to guide running experiment	Day-long experiment shifts, bursty use of NERSC during shift	Minutes	10s-100s nodes	Reservations
Dark Energy Spectroscopic Instrument	Analyze telescope data	Need results by breakfast to guide following night	Deadline in hours	10s of nodes	Realtime
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT	Supernova neutrino burst	Random, no advance notice	Hours	100s of nodes	???

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 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

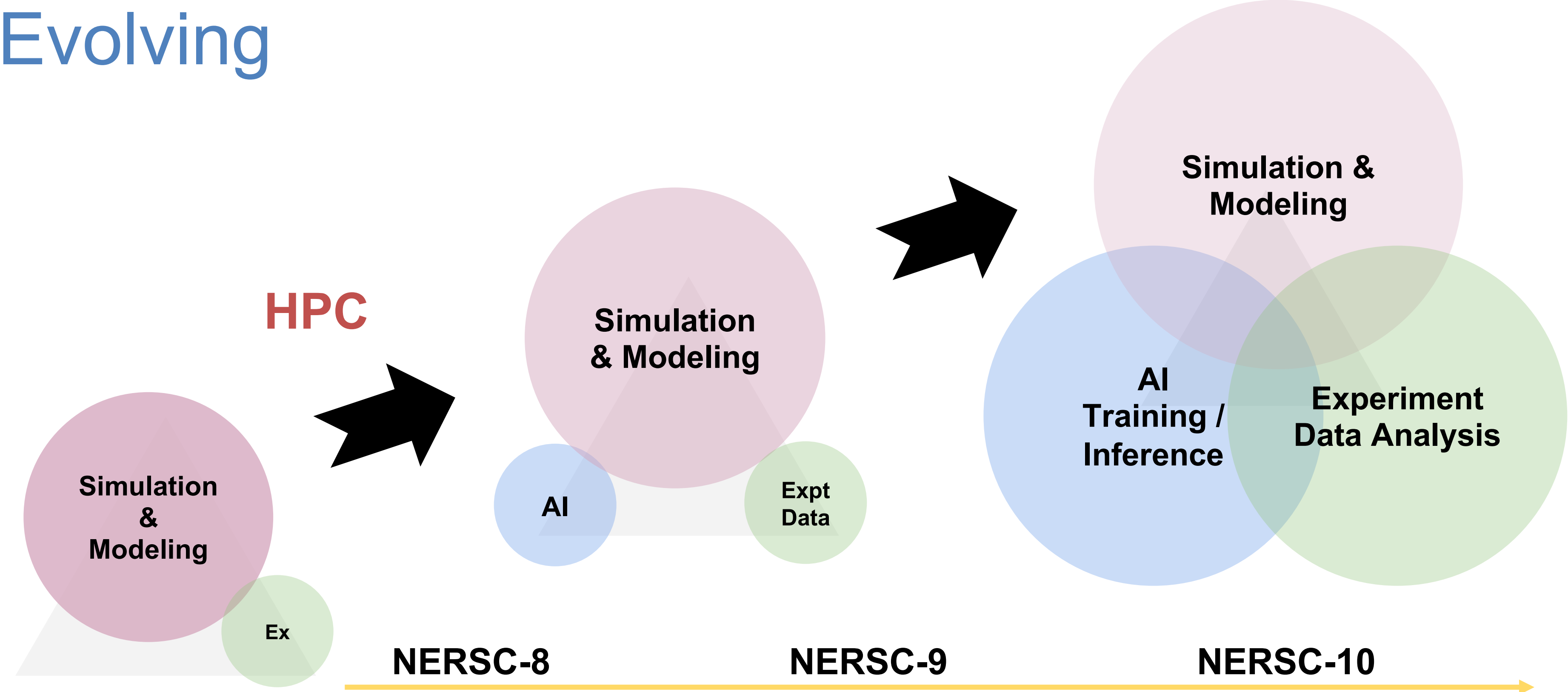
Container build wrappers for multiple architectures, MPI libs etc

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 multiple computing resources***

HPC Facility Workload Balance is Evolving



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.*

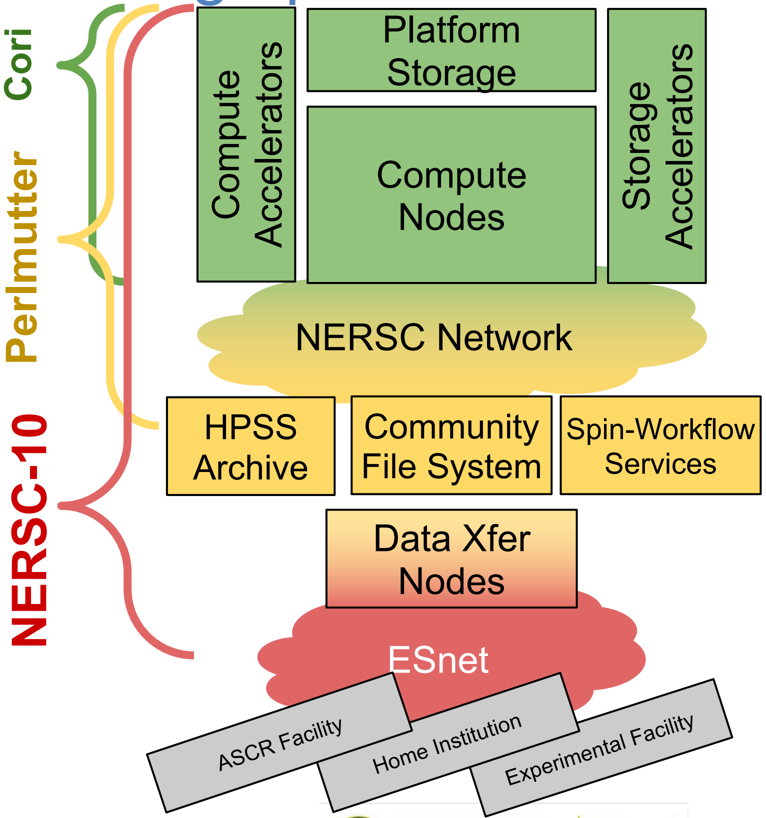
ASCR
ADVANCED SCIENTIFIC COMPUTING RESEARCH
NP
NUCLEAR PHYSICS
STREAM 2016
BASIC RESEARCH NEEDS FOR
Innovation and Discovery of Transformative Experimental Tools
THE FUTURE OF SCIENTIFIC WORKFLOWS
ASCR WORKSHOP ON IN SITU DATA MANAGEMENT
Enabling Scientific Discovery from Diverse Data Sources
January 28–29, 2019
STORAGE SYSTEMS AND I/O: ORGANIZING, STORING, AND ACCESSING DATA FOR SCIENTIFIC DISCOVERY
AI FOR SCIENCE
RICK STEVENS
VALERIE TAYLOR
JEFF NICHOLS
ARTHUR BARNEY MACCABE
KATHERINE YELICK
DAVID BROWN

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

Complexity and heterogeneity managed using complementary technologies

- **Programmable infrastructure**: avoid downfalls of one-size-fits-all, monolithic architecture
- **AI and automation**: sensible selection of default behaviours to reduce complexity for users



Conclusions

- World is changing
  - DOE experimental facilities *need* large scale computer, storage, networking
  - Emerging urgent use cases will fundamentally change the balance of supercomputer workload
- How are we adapting to this?
  - Make sure simulations and experimental analysis can co-exist on our systems
  - Design the system from the ground up for HPC *and* EOD
  - Create opportunities for change by adapting the scheduler
- The scheduler is the heart of how we'll adapt
  - Contributions to open source Slurm
  - Funding large scale changes to benefit Experimental Sciences

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov