Topics and Questions

Three Main Questions of Interest:

- What User Needs would be met by a successful Data and Compute Architecture?
- What scientific data, computing tools, and architectures could NASA implement to support Open-Source Science?
- What concepts might most help rapid transition to an Open-Source Science architecture?

Specific areas of interest include:

- **Ad hoc Work Groups** capable of responding perhaps even within days to an emerging science need
- **Frictionless Access** to science data, compute, and ecosystem, meaning low barriers of entry called for as part of democratizing science
- **Reduced Time-to-science** and attention to the full science production lifecycle
- **Community-specific science** meeting specific performance or functional needs
- **Responsiveness to changes** in Scientific Paradigms and Cultures
National Energy Research Scientific Computing Center (NERSC) is the mission High Performance Computing facility for the DOE Office of Science.

8,000 Users
800 Projects
700 Codes
2,000 NERSC citations per year

Simulations at scale

Data analysis support for DOE’s experimental and observational facilities
Photo Credit: CAMERA
2020 NERSC by the Numbers

7,887 ANNUAL USERS

- 29% Graduate Students
- 20% Postdoctoral Fellows
- 16% Staff Scientists
- 11% University Faculty
- 6% Undergraduate Students
- 6% Professional Staff

>1,800 Refereed Publications Cited NERSC

2020 DOE Office of Science Program Usage Breakdown

- 39% Basic Energy Sciences
- 20% High Energy Physics
- 14% Biological and Environmental Research
- 13% Fusion Energy Sciences
- 12% Nuclear Physics
- 2% Advanced Scientific Computing Research
- <1% Small Business Innovation Research

Data Stored

200 Petabytes
NERSC supports a large number of users and projects from DOE SC’s experimental and observational facilities.

~35% (235) of ERCAP projects self identified as confirming the primary role of the project is to 1) analyze experimental data or; 2) create tools for experimental data analysis or; 3) combine experimental data with simulations and modeling.
NERSC Systems Roadmap

NERSC-7: Edison
2.5 PFs
Multi-core CPU
3MW

NERSC-8: Cori
30PFs
Manycore CPU
4MW

NERSC-9: Perlmutter
~120PFs
CPU and GPU nodes
>5 MW

NERSC-10
ExaSystem
~20MW

2013
2016
2021
2025
What’s new?
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**

Credit: Harrington et al. 2021

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

**AI-reconstructed hydrodynamic fields from approximate N-body simulations**
And a changing business landscape challenges us to assure our requirements are inclusive

- Traditional technology integrator landscape contracting (HPE acquires Cray)
- Other integrators potentially interested, but may have less user/system software expertise (Dell, Penguin)
- Hyperscaler vendors growing share of market (Amazon AWS, Microsoft Azure, Google GPC) with other vendors interested in entering.
- AI, Storage and workflow vendors also have promising technology
Total Unique Nonstaff Users, Data Services/Software

Users of NERSC Data Software and Services, 2021

- Al and Analytics
- Cori Data Features
- Data Management
- Transfer and Access
- Workflows

Category | Service          | 2021 unique users
---|------------------|-------------------
Al and Analytics | Python          | 391
              | TensorFlow      | 321
              | Pytorch         | 185
              | Matlab          | 154
              | IDL             | 148
              | Julia           | 64
              | R (quarter year) | 43
              | Mathematica     | 27
              | R Studio        | 5

Cori Data Features | cori interactive queue | 1,936
                   | haswell interactive queue | 1,546
                   | knl interactive queue | 892
                   | Shifter (half year) | 666
                   | shared queue | 351
                   | Burst Buffer | 108
                   | cori bigmem queue | 67
                   | cori transfer queue | 64
                   | cori realtime queue | 15

Data Management | iSpy             | 1,272
                 | netCDF4py       | 374
                 | ROOT            | 265
                 | MongoDB         | 92
                 | PostgreSQL      | 29
                 | MySQL           | 19

Transfer and Access | NERVT logins   | 2,187
                     | Jupyter         | 1,170
                     | Globus Online    | 1,163
                     | NIMachine        | 568
                     | Superfacility API | 8

Workflows | Dask            | 658
           | GNU parallel    | 166
           | workflow nodes  | 110
           | Fineworks       | 72
           | TaskFarmer      | 46
           | parsl           | 28
           | papermill       | 8
NERSC users require a paradigm shift in the way we design, configure and operate HPC systems

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: 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.

Updated! NERSC-10 CD-0 achieved Sept. 2021
Steering experiments in real-time with simulation & data analysis

15 TB/day

DAQ → Spinning Disk → XRootD → SCRATCH, CFS, HPSS

Requirements for large-scale interactive data analysis and simulations to steer experiments in real-time:

- Ability to surge to large part of HPC system
- High speed data transfer via XRootD+ESnet
- Fast feedback to experiment via real-time data analysis, including AI and simulation
- Ability to orchestrate jobs/parameters and interact with active workflows and results in real-time
- Heterogeneous compute/storage technology to optimize workflow components

Analyze Data, Compare to Simulation, Further Train/Parameterize Models

Tweak Experiment
FES: DIII-D workflow with real-time feedback

DIII-D National Fusion Facility (San Diego, CA)

Event: Anomaly detected at DIII-D experiment

Scientists and operators in control room have ~20 minutes to decide what to do before the next discharge

Real-time HPC capabilities in N10 enable:

- Agile experimental decision making tools
- Optimized use of valuable experimental time
- Enhanced scientific productivity

Fast data transfer, real-time compute

Quickly process data, use AI at scale to locate and analyze anomalies, compare against simulations, and use AI to recommend action.

Real-time results from N10 workflow can help answer:

- What caused the anomaly?
- Adjust experimental parameters?
- Safe to continue?
BES: AI-Enhanced Materials Design Workflow

AI accelerated “traditional” simulation at scale = direct access to experimental regime

Need for refinement of AI potential identified

deploy trained model
restart scale simulation

requires dynamic heterogeneous resources, large scale simulations, accelerators for AI inference and training, advanced scheduling and data management available on integrated platform

high throughput high accuracy first principles simulations to refine AI model
HEP: DUNE Supernova Neutrino Trigger

DUNE: Neutrino beam from Fermilab toward detector at Sanford Underground Research Facility

- Neutrino oscillations & interactions with matter
- Core collapse supernovae

“Nearby” core collapse supernova

- Neutrino oscillations & interactions with matter
- Core collapse supernovae

“Nearby” core collapse supernova

- Neutrino oscillations & interactions with matter
- Core collapse supernovae

Requirements for disruptive (~1 trigger/month), high-value science:

- Compute nodes available on demand, no queue wait
- High-bandwidth ingress to compute nodes on demand
- No phone calls, emails, tickets; just scripted API calls
Real-time HPC from N10 could enable:

- Rapid updates to AI models based on streaming data and real-time simulations on HPC
- More accurate automated decision making
- Optimised sensor placement and calibration

Future-looking!

(Are there other emerging workflows for CESD?)
Superfacility Model
The Superfacility Model: an ecosystem of connected facilities, software and expertise to enable new modes of discovery

Superfacility@ LBNL: NERSC, ESnet and CRD working together to support experimental science

- A model to integrate experimental, computational and networking facilities for reproducible science
- Enabling new discoveries by coupling experimental science with large scale data analysis and simulations
Goal:

By the end of CY 2021, 3 (or more) of our 7 science application engagements will demonstrate automated pipelines that analyze data from remote facilities at large scale, without routine human intervention, using these capabilities:

- **Real-time** computing support
- Dynamic, high-performance **networking**
- Data management and movement tools, incl. **Globus**
- **API**-driven automation
- HPC-scale notebooks via **Jupyter**
- Authentication using **Federated Identity**
- Container-based edge services supported via **Spin**
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**

- **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.
NERSC now supports real-time capabilities

- NERSC has a dedicated pool of real-time nodes for approved projects
- NERSC can also support reservations for experiments and enables pre-emptible jobs to keep utilization high
Spin: Container Services for Science

Many projects need more than HPC. **Spin is a platform for services.**

Users deploy their science gateways, workflow managers, databases, and other network services with Docker containers.

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

Some projects using Spin:

- Track and compare analyses of nightly sky surveys
- Classify and store reusable earth sciences data
- Manage production genomic workflows and data at scale
- Process real-time events for dark matter detection
- Explore materials properties or build simulated materials

NERSC
Jupyter: supercharge interactive supercomputing

NERSC leads in HPC-aware Jupyter deployments
- Patterns and frameworks for connecting Jupyter with HPC
- Data analytics/AI platform in an HPC environment
- Interactive visualization and shareable analysis workflows
- Reproducible science through containerization w/SciData Division

Interactive supercomputing: Jupyter Notebook + HPC Workers
- Launch workers in a short turnaround queue, leveraging our API
- Communicate with distributed analytics clusters (e.g. IPyParallel, Dask)

User quote: “Jupyter notebooks are very important for me: The 3 most important things in life: food, shelter and Jupyter… everything else is optional.”
Jupyter Uptick Continues
Federated Identity (FedID): one identity for many facilities

Users link their home identity to their NERSC account, then use it to log in.

- Simplifies cross-facility workflows
- Users have fewer passwords and login pages
- Home institution manages account lifecycle
- NERSC still manages authorization
- Core technology is established and mature
- **Policy/trust decisions were the bulk of our work**

**Future:** DOE DCDE project is building and promoting similar efforts at other sites
Machine-readable supercomputers: the Superfacility API

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

Endpoints currently deployed:

/meta information about this Superfacility API installation
/status NERSC component system health
/account Get accounting information about the user's projects
/utilities basic file browsing, upload and download of small files to and from NERSC
/storage Transfer files between Globus endpoints.
/compute Run commands and manage batch jobs on NERSC compute
/tasks Get information about your pending or completed tasks
/reservations submit and manage future compute reservations

https://api.nersc.gov/
NESAP Highlights: ExaFEL

- Continuous engagements (Hackathons + ECP) resulted in steady performance gains:
  - Throughput has gone from 2Hz (Edison using openMP) to 1.4kHZ (Perlmutter using Kokkos)
  - Steadily offloading parts of workflow to GPUs

- Workflow scales to size of entire machine
  - MPI ranks share GPUs to hide workflow latency which can’t be parallelized

- Use Perlmutter for realtime data analysis for LCLS running experiments
  - Can analyze an experiment in minutes to give immediate feedback to beamline staff
  - Using Superfacility infrastructure to set up ad control complex workflow across multiple sites
NERSC AI Strategy

- **Deploy** optimized hardware and software systems
  - Currently Perlmutter >6000 A100 GPUs; Work with vendors for optimized AI software
  - >6x increase in usage of DL frameworks since 2018
  - Improve performance, e.g. through benchmarking (e.g. [MLPerf HPC](#))

- **Apply** ML for science using cutting-edge methods
  - “NESAP for Learning” application readiness program with postdocs, early access etc.
  - Other targeted engagements that push model development, scale and performance
  - Leverage lessons learned for all users

- **Empower** through seminars, training and schools
  - E.g. Deep Learning at Scale tutorial at Supercomputing ([SC21 material here](#))
Transformative AI for new science - powered By Perlmutter

FourCastNet
Forecasts global weather at high-resolution. Hybrid data/model parallel @ 4000 GPUs
First deep-learning with skill of numerical weather prediction

Self-supervised sky surveys
Stein et. al. (2021)  [arXiv:2110.00023](https://arxiv.org/abs/2110.00023)
Uncovered thousands of undiscovered strong-lenses

CatalysisDL
Co-developed largest catalysis dataset (OC20);
Graph-parallel NN approaches and NeurIPS 2021 Competition

Unfolding for particle physics
H1 Collaboration: announced at DIS2022
New ML approach extracts new physics insights from old data.
Requires Perlmutter for multiple distributed training runs
NERSC users require a paradigm shift in the way we design, configure and operate HPC systems.

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: 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 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
Reconfigurable storage tailors performance to each workflow's characteristics and needs

NERSC-10 will be programmable to optimize for each workflow

1. User requests hardware resources, connections between them, and data placement
2. System schedules CPU, accelerators, storage, networking, and data movement
3. Same resources are later reconfigured to adapt to new requirements

NERSC-10 will achieve this by embracing technology trends

- Disaggregated, software-defined infrastructure to connect heterogeneous components
- AI and automation to manage
  - complexity of scheduling and operations
  - data movement between reconfigurations
  - complexity for users - sensible defaults

Later that day...

Global file system for everyone
Node-local-like SSD for job 1
High IOPS file system for job 2
Unreserved

Global file system for everyone
Node-local-like SSD for job 3
Extremely resilient object store for job 4
Unreserved
Pools of nodes and bandwidth can be reconfigured to support different SC workflows

Software-defined networking redirects bandwidth to paths that need it

Microservices allow services that utilize bandwidth to scale up/down

One hardware pool configurable to...

- **DTNs** - file transfer from external facilities
- **Routers** - stream data directly to compute
- **Movers** - file transfer between storage tiers
- **Metaschedulers** - dispatch units of work to compute
Exploring the merging of HPC and commercial cloud technologies

2016: Traditional HPC system with a ‘data partition’

2020: Integrates Cloud Technologies to improve availability and flexibility; Innovative network that combines high-performance plus flexibility

2025: NERSC-10: Convergence to harness the power of traditional HPC with the flexibility, configurability and interactivity of cloud to support growing science requirements.
Flexible and dynamic scheduling of compute, storage and bandwidth enables a workflow to reserve various resources at different times.
Expanding NERSC’s role in data management and stewardship

Data Repository Services: enable curation, search, tracking of metadata and dataset versions to enable FAIR Principles for DOE Science Projects

- Provide the requisite infrastructure and services
- Leverage and Integrate Key Technologies for Identity Management, Data Publishing, Data Transfer Services, Data Exploration, etc
- Provide infrastructure to rapidly deploy new services via Spin coming from CS Research Community and DOE Science Projects
Some key questions we are asking ourselves

- NERSC-10 vision is about end-to-end workflows - which impacts the whole data center. How do we do that in a traditional procurement model? How can we integrate new technologies more rapidly?
- How do we securely enable end-to-end workflows?
- How much integration can we manage ourselves vs put to vendor
- What is the role of cloud technologies in HPC?
  - We are having a full day offsite on Sept. 22nd to get our heads around this
- How can we support an order of magnitude more users at NERSC coming from DOE experimental facilities?
  - Major implications regarding who is a user, who can access NERSC and how
- What policies and metrics need adjusting in this new environment?
- How can we recruit and retain the staff we need to accomplish our vision?
Increased emphasis on Integration

- Looking at the transition from NERSC-8 to NERSC-9, we are already doing more technology integration on our largest systems
  - Shasta system software hardening
  - Slingshot hardening
  - Scheduling software development/specifications to developers
  - OpenMP NRE
  - More examples
- We expect the level of integration by NERSC staff to increase for NERSC-10
Our strategy thus far for NERSC-10 era

- Allow technology vendors, (other than prime to respond to an RFP) (we are still figuring out procurement mechanisms to do this)
- Search for partners at other labs, particularly in area of system software
- Leverage community efforts (E4S)
- Do a comprehensive market survey
- Dedicated exploration of new areas (cloud tech, advanced scheduling, system software)
- Increase hiring with a focus on advanced system integration/development skills (applies to many groups)