Accelerating Experimental Workflows on NERSC systems

Katie Antypas
NERSC Division Deputy
Jefferson Lab Seminar
May 15, 2019
NERSC is the mission HPC facility for the DOE Office of Science

7,000 Users
800 Projects
700 Codes
~2000 publications per year

Simulations at scale

Data analysis support for DOE’s experimental and observational facilities

Photo Credit: CAMERA
NERSC supports a large number of users and projects from DOE SC’s experimental and observational facilities.

Approximately 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 Directly Supports Office of Science Priorities

2018 Allocation Breakdown (Hours Millions)

- Advanced Computing: 200; 3%
- Biosciences: 195; 3%
- Small Business: 34; 1%
- BES User Facilities: 232; 3%
- Climate/Environment: 845; 12%
- Nuclear Physics: 850; 12%
- High Energy Physics: 1,142; 17%
- Fusion Energy & Plasma Physics: 950; 14%
- Materials Science: 1,099; 16%
- Chemistry & Geoscience: 1,287; 19%

- NERSC Director's Reserve: 850,000,000; 9%
- ALCC: 750,000,000; 8%
- DOE Science Mission: 6,834,000,000; 74%
Jefferson Lab Users

• 14 users from Jefferson Lab have used over 56M hours thus far 2019
• In addition, NERSC is providing support through our director’s reserve to the Glue-X project

<table>
<thead>
<tr>
<th>Alexander</th>
<th>Austregesilo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nathan</td>
<td>Brei</td>
</tr>
<tr>
<td>Robert</td>
<td>Edwards</td>
</tr>
<tr>
<td>Balint</td>
<td>Joo</td>
</tr>
<tr>
<td>David</td>
<td>Lawrence</td>
</tr>
<tr>
<td>Luka</td>
<td>Leskovec</td>
</tr>
<tr>
<td>Gunn Tae</td>
<td>Park</td>
</tr>
<tr>
<td>David</td>
<td>Richards</td>
</tr>
<tr>
<td>Yves</td>
<td>Roblin</td>
</tr>
<tr>
<td>Rocco</td>
<td>Schiavilla</td>
</tr>
<tr>
<td>Raza</td>
<td>Sufian</td>
</tr>
<tr>
<td>Shaoheng</td>
<td>Wang</td>
</tr>
<tr>
<td>Chip</td>
<td>Watson</td>
</tr>
<tr>
<td>He</td>
<td>Zhang</td>
</tr>
</tbody>
</table>
NERSC Systems Roadmap

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

2013

NERSC-8: Cori
- 30PFs
- Manycore CPU
- 4MW

2016

NERSC-9: Perlmutter
- 3-4x Cori CPU and GPU nodes
- >6 MW

2020

NERSC-10 ExaSystem
- ~20MW

2024
Cori System
Cori: Pre-Exascale System for DOE Science

- Cray XC System - heterogeneous compute architecture
  - 9600 Intel KNL compute nodes, >2000 Intel Haswell nodes
- Cray Aries Interconnect
- NVRAM Burst Buffer, 1.6PB and 1.7TB/sec
- Lustre file system 28 PB of disk, >700 GB/sec I/O
- Investments to support large scale data analysis
  - High bandwidth external connectivity to experimental facilities from compute nodes
  - Virtualization capabilities (Shifter/Docker)
  - More login nodes for managing advanced workflows
  - Support for real time and high-throughput queues
  - Data Analytics Software

- New this year: GPU rack integrated into Cori
NERSC Exascale Scientific Application Program (NESAP)

- Prepare DOE SC users for advanced architectures like Cori and Perlmutter
- Partner closely with 20-40 application teams and apply lessons learned to broad NERSC user community.

Vendor Interactions
Developer Workshops
Leverage community efforts
Engage w/ code teams
Postdoc Program
Early Access To KNL
Dungeon Sessions

Result = 3x Average Code Speedup!
Transition of the entire NERSC workload to advanced architectures

To effectively use Cori KNL, users must exploit parallelism, manage data locality and utilize longer vector units. All features that will be present on exascale era systems.
Users Demonstrate Groundbreaking Science Capability

Large Scale Particle in Cell Plasma Simulations

Stellar Merger Simulations with Task Based Programming

Largest Ever Quantum Circuit Simulation

Largest Ever Defect Calculation from Many Body Perturbation Theory > 10PF

Deep Learning at $1544 \text{ (PF)}$ for Climate and HEP

Celeste: 1st Julia app to achieve 1 PF

Galactos: Solved 3-pt correlation analysis for Cosmology @9.8PF
Particle Collision Data at Scale

- BNL STAR nuclear datasets: PB scale
- Reconstruction processing takes months at BNL computing facility
- With help from NERSC consultants & storage experts, & ESNet networking experts, built highly scalable, fault-tolerant, multi-step data-processing pipeline
- Reconstruction process reduced from months to weeks or days
- Scaled up to 25,600 cores with 98% end-to-end efficiency

A series of collision events at STAR, each with thousands of particle tracks and the signals registered as some of those particles strike various detector components.
NERSC-9: Perlmutter
NERSC-9: A System Optimized for Science

• Cray Shasta System providing 3-4x capability of Cori system
• First NERSC system designed to meet needs of both large scale simulation and data analysis from experimental facilities
  – Includes both NVIDIA GPU-accelerated and AMD CPU-only nodes
  – Cray Slingshot high-performance network will support Terabit rate connections to system
  – Optimized data software stack enabling analytics and ML at scale
    – All-Flash filesystem for I/O acceleration
• Robust readiness program for simulation, data and learning applications and complex workflows
• Delivery in late 2020
From the start NERSC-9 had requirements of simulation and data users in mind

- All Flash file system for workflow acceleration
- Optimized network for data ingest from experimental facilities
- Real-time scheduling capabilities
- Supported analytics stack including latest ML/DL software
- System software supporting rolling upgrades for improved resilience
- Dedicated workflow management and interactive nodes
NERSC-9 will be named after Saul Perlmutter

• Winner of 2011 Nobel Prize in Physics for discovery of the accelerating expansion of the universe.

• Supernova Cosmology Project, lead by Perlmutter, was a pioneer in using NERSC supercomputers combine large scale simulations with experimental data analysis.

Login “saul.nersc.gov”
<table>
<thead>
<tr>
<th>Data Features</th>
<th>Cori experience</th>
<th>N9 enhancements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I/O and Storage</strong></td>
<td>Burst Buffer</td>
<td>All-flash file system: performance with ease of data management</td>
</tr>
<tr>
<td><strong>Analytics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Production stacks</td>
<td>User defined images with Shifter NESAP for data</td>
<td>Optimised analytics libraries and deep learning application benchmarks</td>
</tr>
<tr>
<td>- Analytics libraries</td>
<td>New analytics and ML libraries</td>
<td></td>
</tr>
<tr>
<td>- Machine learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Workflow integration</strong></td>
<td>SchedMD</td>
<td>SLURM co-scheduling Workflow nodes integrated</td>
</tr>
<tr>
<td><strong>Data transfer and streaming</strong></td>
<td>SDN</td>
<td>Slingshot ethernet-based converged fabric</td>
</tr>
</tbody>
</table>

**Data Features**
- Cori experience
- N9 enhancements
GPU Partition added to Cori for NERSC-9

- GPU partition added to Cori to enable users to prepare for Perlmutter system
- 18 nodes each with 8 GPUs
- Software support for both HPC simulations and Machine Learning

GPU cabinets being integrated into Cori Sept. 2018
5 ECP Apps Jointly Selected (Participation Funded by ECP)
20 additional teams selected through Open call for proposals.
Access to Cori GPU rack for application readiness efforts.
Significant *NESAP for Data* App Improvements

**Jonathan Madsen**

*TomoPy* (APS, ALS, etc)

- GPU acceleration of iterative reconstruction algorithms
- New results from first NERSC-9 hack-a-thon w/NVIDIA, >200x speedup!

**Laurie Stephey**

*DESI Spectroscopic Extraction*

- Optimization of Python code on Cori KNL architecture
- Code is 4-7x faster depending on architecture and benchmark
Superfacility Model – Supporting Workflows from Experimental Facilities
Superfacility: 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
On-going Engagements with experimental facilities drive our requirements
Building on past success with ALS

- Real-time analysis of ‘slot-die’ technique for printing organic photovoltaics
- Run experiment on ALS
- Use NERSC for data reduction
- Use OLCF to run simultaneous simulations.
- Real-time analysis of combined results at NERSC

What’s needed?
- Automated calendaring, job submission and steering
- Tracking data across multiple sites
- Algorithm development
Leading the way: LCLS-II


What’s needed?
- Automated job submission and steering
- Seamless data movement via ESnet
- Tracking data across multiple sites
- Integration of bursty jobs into NERSC scheduled workload

Diffraction pattern from LU34
LCLS Experiments using NERSC in Production

- LCLS experiment requires larger computing capability to analyze data in real-time: Partnering with NERSC.
- Detector to Cori rate ~ 5GB/s
- Live analysis for beamline staff
- Use compute reservation on Cori
- Feedback rate is ~ 20 images/sec — allows team to keep up with the experiment

Leading the way: NCEM 4D-Stem

What’s needed?

- Edge device design
- Machine Learning
- Automated job submission and steering
- Data search

FPGA based readout system

Segmented HAADF Detector

100,000 frames/s
Pixilated Detector
Enabling Edge Services with Spin

Challenge
• Workflows often require additional edge services (DBs, APIs, Portals) to achieve their science.

Innovation
• NERSC provides Spin, a multi-tenancy, container-based orchestration system, to support user managed edge services
• NERSC provides the infrastructure, users only concern is to provide their services
• Training and user support were implemented to rapidly on-board projects

Impact and Early Successes
• >70 users have taken training and over 90 services have been deployed in production
• A trained user can bring up a new service in a matter of hours with no staff intervention
Project filesystem replacement

- 75 PB available to users by FY202
- 150 - 300 PB by Perlmutter deployment
Open Research Areas
Many open research areas remain to make the superfacility model successful.

<table>
<thead>
<tr>
<th>Acquire/Transfer</th>
<th>Clean/Filter</th>
<th>Use/Reuse</th>
<th>Publish</th>
<th>Preserve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect from sensors, experiments and move from instrument to center</td>
<td>Deploy edge devices and Organize, annotate, filter, encrypt, compress</td>
<td>Analyze, mine, model, learn, infer, derive, predict</td>
<td>Disseminate &amp; aggregate, using portals, databases</td>
<td>Index, curate, age, track provenance, search, purge</td>
</tr>
</tbody>
</table>
Moving Computation to the Data

- Data velocity increasing
- Data sources increasing
  - Not necessarily co-located with HPC centers
- **Solution**: Custom edge computing devices enable processing before it gets to the HPC Center

On Sensor / Field Deployable Processing

Near Sensor and Real Time Processing
Supporting Data Access and Search

ScienceSearch: Scientific data search

Website: http://sciencesearch.lbl.gov

pyCBIR: Image Search

In conclusion

- We are excited for NERSC-9 and the new data capabilities
- We welcome new partnerships around experimental data and workflows
- Some of our leaders at NERSC to engage with

Debbie Bard
- Group Lead for Data Science Engagement Group at NERSC
- Facility engagement, use cases and

Shane Cannon
- Senior Computing Systems Engineer at NERSC
- Shifter, data transfer

Prabhat
- Group Lead for Data and Analytics Services
- Machine Learning/Deep Learning

Cory Snavely
- Group Lead for Infrastructure Services at NERSC
- SPIN, Identity and Access management
API for Experimental Facilities

Job management: submission, monitoring, retries

Data Movement: Between layers, across facilities

Reservations: HPC, Storage, BW

Publish and Share Data

Software section

Manage Identities

Systems section

IAM Service
<table>
<thead>
<tr>
<th>Capabilities</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Transfer + Access</td>
<td><img src="image" alt="globus online" /> <img src="image" alt="GridFTP" /> <img src="image" alt="jupyter" /> ![<img src="image" alt="" /> <img src="image" alt="python" /> <img src="image" alt="django" /> <img src="image" alt="newt" /></td>
</tr>
<tr>
<td>Workflows</td>
<td><img src="image" alt="FireWorks" /> <img src="image" alt="taskfarmer" /></td>
</tr>
<tr>
<td>Data Management</td>
<td><img src="image" alt="HDF" /> <img src="image" alt="netCDF" /> <img src="image" alt="ROOT" /> <img src="image" alt="mongoDB" /> <img src="image" alt="MySQL" /> <img src="image" alt="PostgreSQL" /></td>
</tr>
<tr>
<td>Data Analytics</td>
<td><img src="image" alt="python" /> <img src="image" alt="Spark" /> <img src="image" alt="TensorFlow" /> <img src="image" alt="PyTorch" /> <img src="image" alt="R" /> <img src="image" alt="julia" /> <img src="image" alt="Caffe" /> <img src="image" alt="K" /></td>
</tr>
<tr>
<td>Data Visualization</td>
<td><img src="image" alt="visit" /> <img src="image" alt="ParaView" /></td>
</tr>
</tbody>
</table>
Identity and Access Management (IAM)

- NERSC is replacing our home grown solution (NIM)!
- New IAM solution will be built with components from Internet2 TIER project
- Benefits for experimental facility workflow users:
  - Simpler account creation
  - Ability for users to have different roles (data users, shell access, web gateway)
  - Transparent and consistent rules for granting access
  - Easier to activate and deactivate accounts, particularly for large projects with many members → better security
  - Native federated identity support!

Identity Enrollment & Registry

Group-based access management
SPIN: Edge Services for Complex Workflows

Container-based platform for easily and quickly creating science gateways, workflow managers and other edge services with limited assistance from staff

- Tightly coupled with HPC resources
- Scalable user defined services
### Elements of the Superfacility model

<table>
<thead>
<tr>
<th>User Engagement</th>
<th>Data Lifecycle</th>
<th>Automated Resource Allocation</th>
<th>Computing at the Edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engage with experimental, observational and distributed sensor user communities to deploy and optimize data pipelines for large-scale systems.</td>
<td>Manage the generation, movement and analysis of data for scalability, efficiency and usability. Enable data reuse and search to increase the impact of experimental, observational and simulation data.</td>
<td>Deliver a framework for seamless resource allocation, calendaring and management of compute, storage and network assets across administrative boundaries.</td>
<td>Design and deploy specialised computing devices for real-time data handling and computation at experimental and computational facilities.</td>
</tr>
</tbody>
</table>
Bringing the Processing to the Data: Edge Computing Embedded Throughout the Workflow

**On Sensor / Field Deployable Processing**

- Can be used in a facility or act as standalone, low-power field deployed unit

**Near Sensor and Real Time Processing**

- Can be used for data reduction, LHC Triggers, or enable new computation, such as a control processor for Quantum Processor

**Smart HPC Interconnects**

**HPC Specialized Accelerators**

Requires expertise across CS Area to provide advances in programming and execution models alongside advanced hardware.