Systems Roadmap and Plans for Supporting Extreme Data Science

Richard Gerber
NERSC Senior Science Advisor

December 10, 2015
Resources for DOE Office of Science Research

Largest funder of physical science research in U.S.

U.S. DEPARTMENT OF ENERGY

Office of Science

Bio Energy, Environment

Computing

Materials, Chemistry, Geophysics

Particle Physics, Astrophysics

Nuclear Physics

Fusion Energy, Plasma Physics
NERSC overview

- Strong focus on science
- 1,808 referred publications in 2014
- Deploy first of a kind systems
- Many users (~6,000)
- Users compute at scale and at high volume
- Diversity of algorithms (~600 codes)
- Extreme scale computing and data analysis
NERSC directly supports DOE’s science mission

- DOE SC offices allocate 80% of the computing and storage resources at NERSC
- ALCC 10%
- NERSC Director’s Reserve 10%
The NERSC-8 System: Cori

- Cori will support the broad Office of Science research community and begin to transition the workload to more energy efficient architectures

- Cray XC system with over 9,300 Intel Knights Landing compute nodes – mid 2016
  - Self-hosted, (not an accelerator) manycore processor with up to 72 cores per node
  - On-package high-bandwidth memory

- Data Intensive Science Support
  - 10 Haswell processor cabinets (Phase 1) to support data intensive applications – Summer 2015
  - NVRAM Burst Buffer to accelerate data intensive applications
  - 28 PB of disk, >700 GB/sec I/O bandwidth

- Robust Application Readiness Plan
  - Outreach and training for user community
  - Application deep dives with Intel and Cray
  - 8 post-docs integrated with key application teams
Intel “Knights Landing” Processor

- Next generation Xeon-Phi, >3TF peak
- Single socket processor - Self-hosted, not a co-processor, not an accelerator
- Up to 72 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread performance over current generation Xeon Phi co-processor (KNC)
- High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory
- Higher performance per watt
- Presents an application porting challenge to efficiently exploit KNL performance features
Cori will be installed in the Computational Research and Theory (CRT) Facility

- Four story, 140,000 GSF, 300 offices, 20Ksf HPC floor, 12.5- >40 MW
- Located for collaboration
  - LBNL, CRD, Esnet, UCB
- Exceptional energy efficiency
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
Cori Phase 1

- Installed in CRT now; running with all NERSC users in pre-production mode
- 1,630 Compute Nodes (52,160 cores)
  - Two Haswell processors/node
  - 16 cores/processor at 2.3 GHz
  - 128 GB DDR4 2133 Mhz memory/ node
- Cray Aries high-speed “dragonfly” topology interconnect
- 22 login nodes for advanced workflows and analytics
- SLURM batch system
- Lustre File system
  - 28 PB capacity, >700 GB/sec peak performance
Cori Phase 1: Nov. 8 – Dec. 8, 2015

83 million NERSC MPP hours delivered to science
NERSC’s Current Big System is Edison

- Edison is the HPCS* demo system (serial #1)
- First Cray Petascale system with Intel processors (Ivy Bridge), Aries interconnect and Dragonfly topology
- Very high memory bandwidth (100 GB/s per node), interconnect bandwidth and bisection bandwidth
- 5,576 nodes, 133K cores, 64 GB/node
- Exceptional application performance

*DARPA High Productivity Computing System program
Edison workload

Since Cori Phase 1 came online Edison has been better serving demand to run large jobs

>16K core jobs using 80% of time now

>32K core jobs using 59% of time

>64K core jobs using 32% of time
## NERSC Systems Timeline

<table>
<thead>
<tr>
<th>Year</th>
<th>System</th>
<th>Model</th>
<th>Processor</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>NERSC-6</td>
<td>Hopper</td>
<td>Cray XE6</td>
<td>1.28 PF</td>
</tr>
<tr>
<td>2014</td>
<td>NERSC-7</td>
<td>Edison</td>
<td>Cray XC30</td>
<td>2.57 PF</td>
</tr>
<tr>
<td>2016</td>
<td>NERSC-8</td>
<td>Cori</td>
<td>Cray XC</td>
<td>30 PF</td>
</tr>
<tr>
<td>2020</td>
<td>NERSC-9</td>
<td></td>
<td></td>
<td>100PF-300PF</td>
</tr>
<tr>
<td>2024</td>
<td>NERSC-10</td>
<td></td>
<td></td>
<td>1EF</td>
</tr>
<tr>
<td>2028</td>
<td>NERSC-11</td>
<td></td>
<td></td>
<td>5-10EF</td>
</tr>
</tbody>
</table>
NERSC 9 Activities

- CD0 signed August 24, 2015
- RFP draft technical specs released Nov. 10, 2015
  - Vendor feedback due Dec. 11, 2015
- Independent Project Review (IPR) Q2CY16
- RFP released late Spring/early Summer 2016
• Goal: Prepare DOE Office of Science user community for Cori manycore architecture
• Partner closely with ~20 application teams and apply lessons learned to broad SC user community
• NESAP activities include:
  - Strong support from vendors
  - Early engagement with code teams
  - Leverage existing community efforts
  - Postdoc Program
  - NERSC training and online modules
  - Early access to KNL technology
  - Developer Workshops for 3rd-Party SW
We are initially focussing on 20 codes

10 codes make up 50% of the workload

25 codes make up 66% of the workload

Edison will be available until 2019/2020

Training and lessons learned will be made available to all application teams
# NESAP Codes

<table>
<thead>
<tr>
<th><strong>Advanced Scientific Computing Research</strong></th>
<th><strong>Basic Energy Sciences</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Almgren (LBNL)</td>
<td>Kent (ORNL)</td>
</tr>
<tr>
<td>Trebotich (LBNL)</td>
<td>Espresso</td>
</tr>
<tr>
<td>Chombo-crunch</td>
<td>Deslippe (NERSC)</td>
</tr>
<tr>
<td></td>
<td>Chelikowsky (UT)</td>
</tr>
<tr>
<td></td>
<td>Bylaska (PNNL)</td>
</tr>
<tr>
<td></td>
<td>Newman (LBNL)</td>
</tr>
<tr>
<td><strong>High Energy Physics</strong></td>
<td><strong>Quantum</strong></td>
</tr>
<tr>
<td>Vay (LBNL)</td>
<td>BerkeleyGW</td>
</tr>
<tr>
<td>IMPACT</td>
<td>PARSEC</td>
</tr>
<tr>
<td>Toussaing (Arizona)</td>
<td>NWChem</td>
</tr>
<tr>
<td>Habib (ANL)</td>
<td>EMGeo</td>
</tr>
<tr>
<td><strong>Nuclear Physics</strong></td>
<td><strong>Biological and Environmental Research</strong></td>
</tr>
<tr>
<td>Maris (Iowa St.)</td>
<td>Smith (ORNL)</td>
</tr>
<tr>
<td>Joo (JLAB)</td>
<td>Yelick (LBNL)</td>
</tr>
<tr>
<td>Christ/Karsch</td>
<td>Ringler (LANL)</td>
</tr>
<tr>
<td>(Columbia/BNL)</td>
<td>Johansen (LBNL)</td>
</tr>
<tr>
<td></td>
<td>Dennis (NCAR)</td>
</tr>
<tr>
<td></td>
<td><strong>MFDn</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Gromacs</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Meraculous</strong></td>
</tr>
<tr>
<td></td>
<td><strong>MPAS-O</strong></td>
</tr>
<tr>
<td></td>
<td><strong>ACME</strong></td>
</tr>
<tr>
<td></td>
<td><strong>CESM</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Fusion Energy Sciences</strong></td>
</tr>
<tr>
<td></td>
<td>Jardin (PPPL)</td>
</tr>
<tr>
<td></td>
<td>Chang (PPPL)</td>
</tr>
<tr>
<td></td>
<td><strong>M3D</strong></td>
</tr>
<tr>
<td></td>
<td><strong>XGC1</strong></td>
</tr>
</tbody>
</table>
To run effectively on Cori users will have to:

• **Manage Domain Parallelism**
  – independent program units; explicit

• **Increase Thread Parallelism**
  – independent execution units within the program; generally explicit

• **Exploit Data Parallelism**
  – Same operation on multiple elements

• **Improve data locality**
  – Cache blocking; Use on-package memory

```plaintext
DO I = 1, N
  R(I) = B(I) + A(I)
ENDDO
```
Resources for Code Teams

- Early access to hardware
  - Access to Babbage (KNC cluster) and early “white box” test systems expected in early 2016
  - Early access and significant time on the full Cori system
- Technical deep dives
  - Access to Cray and Intel staff on-site staff for application optimization and performance analysis
  - Multi-day deep dive (‘dungeon’ session) with Intel staff at Oregon Campus to examine specific optimization issues
- User Training Sessions
  - From NERSC, Cray and Intel staff on OpenMP, vectorization, application profiling
  - Knights Landing architectural briefings from Intel
- NERSC Staff as Code Team Laisons (Hands on assistance)
- Hiring for application performance expertise
- 8 Postdocs
NESAP Timeline

- Jan 2014: Prototype Code Teams (BerkeleyGW / Staff)
  - Prototype good practices for dungeon sessions and use of on site staff.

- May 2014: Requirements Evaluation

- Jan 2015: Gather Early Experiences and Optimization Strategy

- Jan 2016: Vendor General Training

- Jan 2017: Vendor General Training

- Jan 2016: NERSC Led OpenMP and Vectorization Training (One Per Quarter)

- Code Team Activity

- Post-Doc Program

- Center of Excellence

- Chip Vendor On-Site Personnel / Dungeon Sessions

- White Box Access

- Delivery

- NERSC User and 3rd Party Developer Conferences
### NESAP Code Status

**Advanced (waiting for hardware)**
- Chroma
- DWF
- Gromacs
- BerkeleyGW
- MILC
- HACC

**Lots of Progress**
- EMGEO
- Boxlib
- WARP
- XGC1

**Moving**
- PARSEC
- Chombo
- MFDN
- Meraculous
- NWChem

**Need Lots of Work**
- CESM
- ACME
- MPAS
What has gone well

1. Setting requirements for Dungeon Session (Dungeon Session Worksheet).
2. Engagement with IXUG and user communities (DFT, Accelerator Design for Exascale Workshop at CRT)
3. Large number of NERSC and Vendor Training (Vectorization, OpenMP, Tools/Compilers)
   Well Received
4. Learned a Massive Amount about Tools and Architecture
5. Cray COE VERY helpful to work with. Very pro-active.
6. Pipelining Code Work Via Cray and Intel resources

Warp Vectorization Improvements at The Dungeon -
Directly enabled by tiling work with Cray COE

<table>
<thead>
<tr>
<th>Application</th>
<th>All memory on far memory</th>
<th>All memory on near memory</th>
<th>Key arrays on near memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>BerkeleyGW</td>
<td>baseline</td>
<td>52% faster</td>
<td>52.4% faster</td>
</tr>
<tr>
<td>EmGeo</td>
<td>baseline</td>
<td>40% faster</td>
<td>32% faster</td>
</tr>
<tr>
<td>XGC1</td>
<td>baseline</td>
<td>24% faster</td>
<td></td>
</tr>
</tbody>
</table>

Techniques and Tools to Target Arrays for Fastmem:
What Has Gone Well (Continued)

7. Bandwidth sensitive applications that live in HBM expected to perform very well.
8. A lot of Lessons Learned: techniques to place key-arrays in fast-memory, improve prefetching effectiveness, coping without L3 cache etc...
9. CPU Intensive tasks (BGW GPP Kernel) shown to perform well (> Haswell) on early KNL projections if effectively use L2, vectorize well and can make use of 2 VPUs.

The N9 workload analysis shows a large fraction of jobs use < 16GB of memory per node.
NESAP Plans

• Increase excitement and effort in 2016 with extra training events, on-site hackathons and more dungeon sessions with KNL hardware in first 9 months of year (5-6).

• Continue successful Cray+Intel pipelining approach.

• Continue App-Readiness (and post-doc program) as an ongoing center effort through 2025 (exascale).

• Maintain a community database of lessons learned and programming “pearls” for many-core that is searchable by keywords like “vectorization”, “latency”, “stencil” as a standalone portal
Extreme Data Science is Playing a Key Role in Scientific Discovery

Martin Karplus

Saul Perlmutter

George Smoot

Warren Washington
Solving the Puzzle of the Neutrino

- HPC and ESnet vital in the measurement of the important $\theta_{13}$ neutrino parameter.
  - Last and most elusive piece of a longstanding puzzle: why neutrinos appear to vanish as they travel
  - The result affords new understanding of fundamental physics; may eventually help solve the riddle of matter-antimatter asymmetry in the universe.

- HPC for simulation / analysis; HPSS and data transfer capabilities; NGF and Science Gateways for distributing results
  - All the raw, simulated, and derived data are analyzed and archived at a single site
  - => Investment in experimental physics requires investment in HPC.

- One of Science Magazine’s Top-Ten Breakthroughs of 2012

The Daya Bay experiment counts antineutrinos at three detectors (shown in yellow) near the nuclear reactors and calculates how many would reach the detectors if there were no oscillation.

PI: Kam-Biu Luk (LBNL)
Nobel Prize in Physics 2015

Scientific Achievement
The discovery that neutrinos have mass and oscillate between different types

Significance and Impact
The discrepancy between predicted and observed solar neutrinos was a mystery for decades. This discovery overturned the Standard Model interpretation of neutrinos as massless particles and resolved the “solar neutrino problem”

Research Details
The Sundbury Neutrino Observatory (SNO) detected all three types (flavors) of neutrinos and showed that when all three were considered, the total flux was in line with predictions. This, together with results from the Super Kamiokande experiment, was proof that neutrinos were oscillating between flavors and therefore had mass

NERSC helped the SNO team use PDSF for critical analysis contributing to their seminal PRL paper. HPSS serves as a repository for the entire 26 TB data set.


Nobel Recipients: Arthur B. McDonald, Queen’s University (SNO) Takaaki Kajita, Tokyo University (Super Kamiokande)
NERSC has been supporting data intensive science for a long time

- Palomar Transient Factory Supernova
- Planck Satellite Cosmic Microwave Background Radiation
- Alice Large Hadron Collider
- Atlas Large Hadron Collider
- Dayabay Neutrinos
- ALS Light Source
- LCLS Light Source
- Joint Genome Institute Bioinformatics
NERSC users import more data than they export!

Importing more than 1PB/month

Exporting more than 1PB/month
NERSC archives an enormous amount of data for the scientific community.

**Archive Data Breakdown**

- ClimateResearch: 31%
- Astrophysics: 17%
- ComputerScience: 11%
- Biosciences: 10%
- NuclearPhysics: 6%
- MaterialsScience: 5%
- LatticeQCD: 4%
- AppliedMath: 4%
- FusionEnergy: 2%
- HEP: 2%
- Combustion: 2%
- Other: 6%

60 PB of data are stored in NERSC’s HPSS Archive.
NERSC’s Goal for Data Initiative

Increase the productivity, usability, and impact of DOE’s experimental user facilities and other data-intensive science by providing comprehensive data systems and services to store, analyze, manage, and share data.
Through 2015 NERSC deployed separate Compute Intensive and Data Intensive Systems

**Compute Intensive**

**Data Intensive**

Carver

Genepool

PDSF
But how different really are the compute and data intensive platforms?

<table>
<thead>
<tr>
<th>Policies</th>
<th>Software/Configuration</th>
<th>Hardware</th>
</tr>
</thead>
</table>
| • Fast-turn around time. Jobs start shortly after submitted  
• Can run large numbers of throughput jobs | • Support for complex workflows  
• Communication and streaming data from external databases and data sources  
• Easy to customize user environment | • Local disk for fast I/O  
• Some systems (not all) have larger memory nodes  
• Support for advanced workflows (DB, web, etc) |

Differences are primarily software and policy issues with some hardware differences in the ratio of I/O, memory and compute
NERSC is making significant investments on Cori to support data intensive science

- New queue policies: real time, and high throughput queues
- High bandwidth external connectivity to databases from compute nodes
- More (23) login nodes for managing advanced workflows
- Virtualization capabilities (Docker)
- NVRAM Flash Burst Buffer as I/O accelerator
  - 1.5PB, 1.5 TB/sec
  - User can request I/O bandwidth and capacity at job launch time
  - Use cases include, out-of-core simulations, image processing, shared library applications, heavy read/write I/O applications
Burst Buffer Motivation

- Flash storage is significantly more cost effective at providing bandwidth than disk (up to 6x)
- Flash storage has better random access characteristics than disk, which help many SC workloads
- Users’ biggest request (complaint) after wanting more cycles, is for better I/O performance

Application perceived I/O rates, with no burst buffer (top), burst buffer (bottom).

Analysis from Chris Carothers (RPI) and Rob Ross (ANL)
Burst Buffer Use Cases

• Accelerate I/O
  – Checkpoint/restart or other high bandwidth reads/writes
  – Apps with high IOP/s e.g. non-sequential table lookup
  – Out-of-core applications
  – Fast reads for image analysis

• Advanced Workflows
  – Coupling applications, using the Burst Buffer as interim storage
  – Streaming data from experimental facilities

• Analysis and Visualization
  – In-situ/ in-transit
  – Interactive visualization
Burst Buffer Software Development Efforts

- Scheduler enhancements
  - Automatic migration of data to/from flash
  - Dedicated provisioning of flash resources
  - Persistent reservations of flash storage
- Caching mode – data transparently captured by the BB nodes
  - Transparent to user -> no code modifications required
- Enable In-transit analysis
  - Data processing or filtering on the BB nodes – model for exascale

Create Software to enhance usability and to meet the needs of all NERSC users
Burst Buffer Early User Program call for proposals

• Aug 10th: solicited proposals for BB Early Users program.
  – Award of exclusive early use of BB on Cori P1, plus help of NERSC experts to optimise application for BB.

• Selection criteria include:
  – Scientific merit.
  – Computational challenges.
  – Cover range of BB data features.
  – Cover range of DoE Science Offices.

• Great interest from the community, 29 proposals received. Good distribution across offices...
Many great applications...

- We’re very happy with the response to the program call.
- Decided to support more applications than we’d originally anticipated
- Other applications will not be supported by NERSC staff, but will have early access to Cori P1 and the BB.
- Breakdown by DoE Office:

<table>
<thead>
<tr>
<th></th>
<th>ASCR</th>
<th>BER</th>
<th>BES</th>
<th>Fusion</th>
<th>HEP</th>
<th>Nuclear</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERSC Supported</td>
<td>1.5</td>
<td>2.5</td>
<td>2.5</td>
<td>1</td>
<td>4.5</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Early Access</td>
<td>3</td>
<td>7</td>
<td>2.5</td>
<td>0</td>
<td>2.5</td>
<td>0</td>
<td>15</td>
</tr>
</tbody>
</table>
## NERSC supported projects

<table>
<thead>
<tr>
<th>Project</th>
<th>DoE office</th>
<th>BB data features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nyx/Boxlib cosmology simulations <em>(Ann Almgren, LBNL)</em></td>
<td>HEP</td>
<td>I/O bandwidth with BB; checkpointing; workflow application coupling; in-situ analysis.</td>
</tr>
<tr>
<td>Phoenix: 3D atmosphere simulator for supernovae <em>(Eddie Baron, U. Oklahoma)</em></td>
<td>HEP</td>
<td>I/O bandwidth with BB; staging intermediate files; workflow application coupling; checkpointing.</td>
</tr>
<tr>
<td>Chombo-Crunch + Visit for carbon sequestration <em>(David Trebotich, LBNL)</em></td>
<td>BES</td>
<td>I/O bandwidth with BB; in-situ analysis/visualization using BB; workflow application coupling.</td>
</tr>
<tr>
<td>Sigma/UniFam/Sipros Bioinformatics codes <em>(Chongle Pan, ORNL)</em></td>
<td>BER</td>
<td>Staging intermediate files; high IOPs; checkpointing; fast reads.</td>
</tr>
<tr>
<td>XGC1 for plasma simulation <em>(Scott Klasky, ORNL)</em></td>
<td>Fusion</td>
<td>I/O bandwidth with BB; intermediate file I/O; checkpointing.</td>
</tr>
<tr>
<td>PSANA for LCLS <em>(Amadeo Perazzo, SLAC)</em></td>
<td>BES/BER</td>
<td>Staging data with BB; workflow management; in-transit analysis.</td>
</tr>
</tbody>
</table>
## NERSC supported projects

<table>
<thead>
<tr>
<th>Project</th>
<th>DoE office</th>
<th>BB data features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALICE data analysis <em>(Jeff Porter, LBNL)</em></td>
<td>NP</td>
<td>I/O bandwidth with BB; read-intensive I/O.</td>
</tr>
<tr>
<td>Tractor: cosmological data analysis <em>(DESI)</em> <em>(Peter Nugent, LBNL)</em></td>
<td>HEP</td>
<td>Intermediate file I/O using BB; high IOPs.</td>
</tr>
<tr>
<td>VPIC-IO performance <em>(Suren Byna, LBNL)</em></td>
<td>HEP/ACSR</td>
<td>I/O bandwidth with BB; in-situ data analysis; BB to stage data.</td>
</tr>
<tr>
<td>YODA: Geant4 sims for ATLAS detector <em>(Vakhtang Tsulaia, LBNL)</em></td>
<td>HEP</td>
<td>BB for high IOPs; stage small intermediate files.</td>
</tr>
<tr>
<td>ALS SPOT Suite <em>(Craig Tull, LBNL)</em></td>
<td>BES/BER</td>
<td>BB as fast cache; workflow management; visualization.</td>
</tr>
<tr>
<td>TomoPy for ALS image reconstruction <em>(Craig Tull, LBNL)</em></td>
<td>BES/BER</td>
<td>I/O throughput with BB; workflow management; read-intensive I/O.</td>
</tr>
<tr>
<td>kitware: VPIC/Catalyst/ParaView <em>(Berk Geveci, kitware)</em></td>
<td>ASCR</td>
<td>in-situ analysis/visualization with BB; multi-stage workflow.</td>
</tr>
</tbody>
</table>
A variety of use cases are represented by the BB Early Users

<table>
<thead>
<tr>
<th>Application</th>
<th>I/O bandwidth: reads</th>
<th>I/O bandwidth: writes (checkpointing)</th>
<th>High IOPs</th>
<th>Workflow coupling</th>
<th>In-situ / in-transit analysis and visualization</th>
<th>Staging intermediate files/pre-loading data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nyx/Boxlib</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Phoenix 3D</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Chomo/Crunch + Visit</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Sigma/UniFam/Sipros</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>XGC1</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PSANA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ALICE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Tractor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>VPIC/IO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>YODA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ALS SPOT/TomoPy</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>kitware</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
We’ve heard from a number of users that the lack of ‘realtime’ access to the system is a barrier to scientific productivity.

With NERSC’s new batch scheduler, SLURM, we have the capability to offer ‘immediate’ or ‘real-time’ access on Cori Phase 1, for projects and users with requirements for fast turn around.

But how to select and judge projects?

We added a question to ERCAP about realtime needs to assess demand and size realtime resources.

Review with program managers and get feedback.
Immediate Queue – ERCAP Requests

• 19 responses (out of > 700) a small fraction of our workload

• Responses from 5 of 6 Offices, SBIR and EERE, demonstrating need is not confined to one scientific domain

• Expected responses:
  – ALS, Palomar Transient Factory, CRD workflow research, MyGreenCar, OpenMSI, KBASE, Materials analysis, 2 PDSF projects

• A few surprises
  – 3 similar Fusion responses (MIT, GA, LLNL) noting DIII-D (tokomak fusion reactor run by General Atomics) can be adjusted by real-time codes
  – Industry response, Vertum partners, Predictive Power Grid performance, run simulations daily 12 hours apart.
Shifter brings user defined images to supercomputers

• Shifter, a container for HPC, allows users to bring a customized OS environment and software stack to an HPC system.

• Use cases
  – High energy physics collaborations that require validated software stacks
  – Cosmology and bioinformatics applications with many 3rd party dependencies
  – Light source applications that with complicated software stacks that need to run at multiple sites
Upgrading Cori’s External Connectivity

Enable 100Gb+ Instrument to Cori

• Streaming data to the supercomputer allows for analytics on data in motion

• Cori network upgrade provides SDN (software defined networking) interface to ESnet. 8 x 40Gb/s bandwidth.

• Integration of data transfer and compute enables workflow automation

Cori Network Upgrade Use Case:

• X-ray data sets stream from detector directly to Cori compute nodes, removing need to stage data for analysis.

• Software Defined Networking allows planning bandwidth around experiment run-time schedules

• 150TB bursts now, LCLS-II has 100x data rates
Superfacility Concept
New technology innovations are challenging existing ways of scientific discovery

- Data volumes are increasing faster than Moore’s Law
- Facility data exceeds local computing and networking capabilities
- Unfeasible to put a supercomputing center at every experimental facility
Fortunately we have ESnet to connect all our facilities.
Based on our experiences supporting science from experimental facilities, we see a Common Design Pattern

- **predictable data movement**
- **Local Data processing/filtering**
- **Data generation from Sensors**
- **on-the-fly calibration**
- **Analysis and modeling**
- **Real-time access and Visualization**
- **Storage, Archive and Share**
- **On-site Scientist**
- **Remote users**
Superfacility Vision: A network of connected facilities, software and expertise to enable new modes of discovery

- New mathematical analyses
- Integrated with ESnet: Designed for Big Science Data
- Experimental Facilities
- Fast implementations on latest computers
- Real-time analysis and data management
- Computing Facilities
NERSC has engagements with a number of experimental facilities

- ALS – Advance Light Source
- LCLS – Linac Coherent Light Source
- LHC through PDSF collaboration
- JGI – Joint Genome Institute
Thank you!