Overview of Cori

Sudip Dosanjh
Director

November 21, 2014
NERSC’s latest system is Edison

- Edison is a HPCS demo system (serial #1)
- First Cray Petascale system with Intel processors, Aries interconnect and Dragonfly topology
- Very high memory bandwidth (100 GB/s per node), interconnect bandwidth and bisection bandwidth
- 64 GB/node
- Exceptional application performance
• Edison doesn’t deploy accelerators or GPUs
• Disruptions in programming models are a challenge for NERSC
  – Many users
  – Many codes
  – We don’t select our users
We support a broad user base

- 5000 users, and we typically add 350 per year
- Geographically distributed: 47 states as well as multinational projects
NERSC Users at ASCAC Institutions

- Boston University – 2
- New York University - 4
- University of Houston – 26
- Google Inc. – 0
- Sandia National Lab – 69
- University of Tennessee – 50
- University of Michigan -60
- Microsoft Research – 0
- Pittsburgh Supercomputer Center – 0 (CMU: 15 - U Pittsburgh: 3)
- MIT– 120
- UC Merced – 4
- UC Santa Barbara – 39
- Rice University – 33
- LLNL– 77
- Total: 467
We support a diverse workload

- Many codes (600+) and algorithms
- Computing at scale and at high volume

---

![Job Size Breakdown on Edison](image)

- 65,536+ cores
- 16,384-65,535 cores
- 8,192-16,383 cores
- 1,024-8,191 cores
- 1-1,023 cores
We directly support DOE’s science mission

• We are the primary computing facility for DOE Office of Science

• DOE SC allocates the vast majority of the computing and storage resources at NERSC
  – Six program offices allocate their base allocations and they submit proposals for overtargets
  – Deputy Director of Science prioritizes overtarget requests

• Usage shifts as DOE priorities change
What’s changed for Cori?

• Heightened awareness among application teams
• Many codes are being adapted for next generation systems
• Technology changes (e.g., self-hosted many core chips, tighter CPU/GPU integration) will make the transition easier
• We must transition to energy efficient architectures to meet the science needs of our users
Keeping up with user needs will be a challenge
NERSC Supports Science Needs at Many Difference Scales and Sizes

High Throughput: Statistics, Systematics, Analysis, UQ

Larger Physical Systems, Higher Fidelity

<table>
<thead>
<tr>
<th>Core-Hours Used (Millions)</th>
<th>1-31</th>
<th>32-63</th>
<th>64-511</th>
<th>512-1K</th>
<th>1K-4K</th>
<th>4K-8K</th>
<th>8K-16K</th>
<th>16K-32K</th>
<th>32K-64K</th>
<th>65-128K</th>
<th>128K+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>675,233</td>
<td>30,991</td>
<td>123,776</td>
<td>15,672</td>
<td>14,618</td>
<td>1,604</td>
<td>826</td>
<td>667</td>
<td>202</td>
<td>56</td>
<td>15</td>
</tr>
<tr>
<td>Core-Hours</td>
<td>7</td>
<td>10.7</td>
<td>55</td>
<td>31</td>
<td>49</td>
<td>28</td>
<td>7</td>
<td>66</td>
<td>30</td>
<td>8.7</td>
<td>0.382</td>
</tr>
</tbody>
</table>

- 10 -
NERSC needs to transition to energy efficient architectures

Manycore or Hybrid is the only approach that crosses the exascale finish line
NERSC-8 (Cori) Mission Need

The Department of Energy Office of Science requires an HPC system to support the rapidly increasing computational demands of the entire spectrum of DOE SC computational research.

• Provide a significant increase in computational capabilities, at least 10 times the sustained performance of the Hopper system on a set of representative DOE benchmarks

• Delivery in the 2015/2016 time frame

• Provide high bandwidth access to existing data stored by continuing research projects.

• Platform needs to begin to transition users to more energy-efficient many-core architectures.
ACES and NERSC formed a partnership for next-generation supercomputers

- Visible collaboration between ASCR and ASC
- Strengthen impact on industry
- Address challenges transitioning applications to advanced manycore architectures with a broader coalition
- Act as a risk mitigation strategy for NERSC-8 and Trinity systems by having a partner to work with on technical challenges deploying and testing NERSC-8 and Trinity

Alliance for application Performance at the EXtreme scale (APEX)
This was a collaboration of two separate projects

Joint

NERSC-8 Mission Drivers

Market surveys

Creating requirements

Release RFP

Vendor Selection

Negotiations

Separate Contracts

Trinity System

NERSC-8 System

Trinity Mission Drivers

Joint Quarterly Reviews, bug reports, collaboration on application transition, advanced options
The Cori system
Cori Configuration

• **64 Cabinets of Cray XC System**
  – Over 9,300 ‘Knights Landing’ compute nodes
    • 64-128 GB memory per node
  – Over 1900 ‘Haswell’ compute nodes
    • Data partition
  – 14 external login nodes
  – Aries Interconnect (same as on Edison)
  – > 10x Hopper sustained performance using NERSC SSP metric

• **Lustre File system**
  – 28 PB capacity, 432 GB/sec peak performance

• **NVRAM “Burst Buffer” for I/O acceleration**

• **Significant Intel and Cray application transition support**

• **Delivery in mid-2016; installation in new LBNL CRT**
Intel “Knights Landing” Processor

- Next generation Xeon-Phi, >3TF peak
- Single socket processor - Self-hosted, not a co-processor, not an accelerator
- Greater than 60 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™
- Intel® "Silvermont" architecture enhanced for high performance computing
- 512b vector units (32 flops/clock – AVX 512)
- 3X single-thread performance over current generation Xeon-Phi co-processor
- High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory
- Higher performance per watt
Cori will be installed in the Computational Research and Theory (CRT) Facility

- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 40 MW to building

- **Located for collaboration**
  - CRD and ESnet
  - UC Berkeley

- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
  - LEED gold design

- **Initial occupancy early 2015**
Application Readiness -- Challenges and Strategy
We will initially focus 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
We are launching the NERSC Exascale Science Applications Program (NESAP)

• NESAP components:
## 20 NESAP Tier-1 and Tier-2 codes

<table>
<thead>
<tr>
<th>Advanced Scientific Computing Research</th>
<th>Basic Energy Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almgren (LBL) – BoxLib AMR Framework</td>
<td>Kent (ORNL) – Quantum Espresso</td>
</tr>
<tr>
<td>Trebotich (LBL) – Chombo-crunch</td>
<td>Deslippe (NERSC) – BerkeleyGW</td>
</tr>
<tr>
<td></td>
<td>Chelikowsky (UT) – PARSEC</td>
</tr>
<tr>
<td></td>
<td>Bylaska (PNNL) – NWChem</td>
</tr>
<tr>
<td></td>
<td>Newman (LBL) – EMGeo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High Energy Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vay (LBL) – WARP &amp; IMPACT</td>
</tr>
<tr>
<td>Toussaint(Arizona) – MILC</td>
</tr>
<tr>
<td>Habib (ANL) – HACC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nuclear Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maris (Iowa St.) – MFDn</td>
</tr>
<tr>
<td>Joo (JLAB) – Chroma</td>
</tr>
<tr>
<td>Christ/Karsch (Columbia/BNL) – DWF/HISQ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biological and Environmental Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith (ORNL) – Gromacs</td>
</tr>
<tr>
<td>Yelick (LBL) – Meraculous</td>
</tr>
<tr>
<td>Ringler (LANL) – MPAS-O</td>
</tr>
<tr>
<td>Johansen (LBL) – ACME</td>
</tr>
<tr>
<td>Dennis (NCAR) – CESM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fusion Energy Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jardin (PPPL) – M3D</td>
</tr>
<tr>
<td>Chang (PPPL) – XGC1</td>
</tr>
</tbody>
</table>
Comparison of Selected Apps with 2013 Usage

Breakdown of Application Hours on Hopper and Edison 2013

Tier-1, 2 Code
Tier-3
Tier-1, 2 Proxy Code
Other Codes
The Selected Codes are Diverse in Several Dimensions

Almost all others

production codes vs. up and coming

MPAS-O
ACME
PARSEC

Higher vs. lower readiness

HACC, Chroma, MILC, BerkeleyGW, XGC

M3D, CESM, EMgeo, WARP & Synergia, Crunch

Smaller community apps vs Overlap with established ALCF/OLCF readiness codes

EMGeo, WARP, Meraculous, Chombo-crunch, BoxLib

MILC, CESM, Chroma, HISQ/DWF, Gromacs, QE, NWChem
NESAP has already received recognition
NESAP Intel and Cray collaboration improved BerkeleyGW performance

NESAP (Cray COE and Intel Dungeon Session) advances for BerkeleyGW kernels include:
1. Thread scaling improvements beyond 10 threads
2. Addition of cache-blocking in bandwidth-bound kernels
3. Improved vectorization in kernels (including small matmuls)
Extreme Data Science
DOE Facilities are Facing a Data Deluge

- Astronomy
- Genomics
- Climate
- Physics
- Light Sources
Exponentially increasing data traffic

NERSC daily routed WAN traffic since 2002

- First petabyte day expected in 2020
- Jump driven by data intensive applications
- Major improvements in TCP auto-tuning
NERSC users import more data than they export!

- Importing more than 1PB/month
- Exporting more than 1PB/month
Extreme Data Science is Playing a Key Role in Scientific Discovery

- **Measurement of the important** $\theta_{13}$ **neutrino parameter. One of Science Magazine’s Top-Ten Breakthroughs of 2012.**
  - Last and most elusive piece of a longstanding puzzle: why neutrinos appear to vanish as they travel

- The Palomar Transient Factory Discovered over 2000 supernovae in the last 5 years, including the youngest and closest Type Ia supernova in past 40 years

- Trillions of measurements by the Planck satellite led to the most detailed maps ever of cosmic microwave background (One of Physics Today’s Top 10 breakthroughs of 2013)

- Materials project has over 5000 users and was featured on the cover of Scientific American

---

**SN 2011fe**

Ph. Shri Kulkarni (Caltech)
We currently deploy separate Compute Intensive and Data Intensive Systems

**Compute Intensive**

Edison

Hopper

**Data Intensive**

Carver

Genepool

PDSF
The Need for Data Intensive Systems

- Communicate with databases / host databases
- Complex workflows (including High Throughput Computing - HTC)
- Policy flexibility
- Local disk
- Very large memory
- Massive serial jobs (~100K)
- Easy to customize environment and the environment is familiar

- Dramatically growing data sets require Petascale+ computing for analysis
- In addition, we increasingly need to couple large-scale simulations and data analysis
Baryon Acoustic Oscillations (BAO):

Large quantities of data need to be analyzed.

- Imaging survey in 2005: 20 TB
- in 2025: 60 PB

Statistical analyses need MCMC for cross-correlation of the millions of galaxies
-- collapsing the problem to just 2-point statistics.

All data analysis dependent on comparisons to supercomputer-based N-body simulations of the evolution of matter in the universe.

- Current state of art: $2048^3 - 4096^3$ “particles.”
- Need an order of magnitude more.
Cosmic Microwave Background (CMB):
Exponentially growing data chasing fainter echos:
- BOOMERanG: $10^9$ samples in 2000
- Planck: $10^{12}$ samples in 2013 (0.5 PB)
- CMBpol: $10^{15}$ samples in 2025

Uncertainty quantification through Monte Carlos
- Simulate $10^4$ realizations of the entire mission
- Control both systematics and statistics

Mission-class science relies on HPC evolution.
Cori Data Enhancements

- Data partition with large memory nodes and throughput optimized processors
- Burst buffer -- NVRAM nodes on the interconnect fabric for IO caching
- Larger disk system

Goals are to enable the analysis of large experimental data sets and in-situ analysis coupled to Petascale simulations
## Popular features of a data intensive system can be supported on Cori

<table>
<thead>
<tr>
<th>Data Intensive Workload Need</th>
<th>Cori Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Disk</td>
<td>NVRAM ‘burst buffer’</td>
</tr>
<tr>
<td>Large memory nodes</td>
<td>128 GB/node on Haswell; Option to purchase fat (1TB) login node</td>
</tr>
<tr>
<td>Massive serial jobs</td>
<td>NERSC serial queue prototype on Edison; MAMU</td>
</tr>
<tr>
<td>Complex workflows</td>
<td>More (14) external login nodes; CCM mode for now</td>
</tr>
<tr>
<td>Communicate with databases from compute nodes</td>
<td><strong>Compute Gateway Node</strong></td>
</tr>
<tr>
<td>Stream Data from observational facilities</td>
<td><strong>Compute Gateway Node</strong></td>
</tr>
<tr>
<td>Easy to customize environment</td>
<td><strong>User Defined Images</strong></td>
</tr>
<tr>
<td>Policy Flexibility</td>
<td>Improvements coming with Cori: Rolling upgrades, CCM, MAMU, above COEs would also contribute</td>
</tr>
</tbody>
</table>
Conclusions
The Cori System

Cray XC40-LC 64 Cabinets, Aries Network

HASWELL COMPUTE (1,600+ NODES)

KNL COMPUTE (9,300+ NODES, 60+ CORES EACH, 64-128GB RAM, 16GB HIGH-BW MEMORY)

BURST BUFFER (1.5+ PB capacity, >1.5TB/sec B/W)

System Management
Compute Network Nodes
I/O Network Nodes (Internal & External)
Workload Management Support Nodes
Dynamic Shared Library Support Nodes
External IB Network
GigE Network
40 GE Network
10 GE Network
Storage System:
Parallel Filesystem (Lustre)
28+ PB
430+ GB/sec
Our goal is to enable science that can’t be done on today’s supercomputers

Martin Karplus

Saul Perlmutter

George Smoot

Warren Washington