NERSC’s 10 year plan

Sudip Dosanjh
Director

February 5, 2014
NERSC’s 40th Anniversary!

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1974</td>
<td>Founded at Livermore to support fusion research with a CDC system</td>
</tr>
<tr>
<td>1978</td>
<td>Cray 1 installed</td>
</tr>
<tr>
<td>1983</td>
<td>Expanded to support today’s DOE Office of Science</td>
</tr>
<tr>
<td>1986</td>
<td>ESnet established at NERSC</td>
</tr>
<tr>
<td>1994</td>
<td>Cray T3D MPP testbed</td>
</tr>
<tr>
<td>1994-2000</td>
<td>Transitioned users from vector processing to MPP</td>
</tr>
<tr>
<td>1996</td>
<td>Moved to Berkeley Lab</td>
</tr>
<tr>
<td>1996</td>
<td>PDSF data intensive computing system for nuclear and high energy physics</td>
</tr>
<tr>
<td>1999</td>
<td>HPSS becomes mass storage platform</td>
</tr>
<tr>
<td>2006</td>
<td>Facility wide filesystem</td>
</tr>
<tr>
<td>2010</td>
<td>Collaboration with JGI</td>
</tr>
</tbody>
</table>
NERSC collaborates with computer companies to deploy advanced HPC and data resources

- Hopper (N6) and Cielo (ACES) were the first Cray petascale systems with a Gemini interconnect
- Edison (N7) is the first Cray petascale system with Intel processors, Aries interconnect and Dragonfly topology (serial #1)
- N8 and Trinity (ACES) are being jointly designed as on-ramps to exascale
- Architected and deployed data platforms including the largest DOE system focused on genomics
- One of the first facility-wide filesystems

We employ experts in high performance computing, computer systems engineering, data, storage and networking
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 overtargerts
  - Deputy Director of Science prioritizes overtarget requests
- Usage shifts as DOE priorities change
We focus on the scientific impact of our users

• 1,500 journal publications per year
• More than 10 journal cover stories per year
• 3 recent Nobel Prize-winning projects used NERSC (2007, 2011, 2013)
• Physics Magazine 2013 “Breakthrough of the Year” used NERSC resources to identify first high-energy cosmic neutrinos. (IceCube)
• Finding that Earth-like planets are not uncommon in our galaxy recognized as a top 2013 discovery by Wired Magazine and covered in The New York Times.
• MIT researchers developed a new approach for desalinating sea water using sheets of graphene, a one-atom-thick form of the element carbon. Smithsonian Magazine’s fifth “Surprising Scientific Milestone of 2012.”
• Four of Science Magazine’s insights of the last decade (three in genomics, one related to cosmic microwave background)
We support a broad user base

- 4500 users, and we typically add 350 per year
- Geographically distributed: 47 states as well as multinational projects
We support a diverse workload

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

2012 Job Size Breakdown on Hopper

- 65,536+ cores
- 16,384-65,535 cores
- 8,192-16,383 cores
- 1,024-8,191 cores
- 1-1,023 cores
Our operational priority is providing highly available HPC resources backed by exceptional user support

• We maintain a very high availability of resources (>90%)
  – One large HPC system is available at all times to run large-scale simulations and solve high throughput problems

• Our goal is to maximize the productivity of our users
  – One-on-one consulting
  – Training (e.g., webinars)
  – Extensive use of web pages
  – We solve or have a path to solve 80% of user tickets within three business days
NERSC Systems Today

**Edison: 2.39PF, 333 TB RAM**
- Cray XC30
- 5,192 nodes, 125K Cores

**Hopper: 1.3PF, 212 TB RAM**
- Cray XE6
- 6,384 nodes, 150K Cores

**Production Clusters**
- Carver, PDSF, JGI, KBASE, HEP
- 14x QDR

**Vis & Analytics**
- Data Transfer Nodes
- Adv. Arch. Testbeds
- Science Gateways

**Ethernet & IB Fabric**
- Science Friendly Security
- Production Monitoring
- Power Efficiency
- WAN

**Global Scratch**
- 3.6 PB
- 5 x SFA12KE

**/project**
- 5 PB
- DDN9900 & NexSAN

**/home**
- 250 TB
- NetApp 5460

**HPSS**
- 50 PB stored, 240 PB capacity, 20 years of community data

**WAN**
- 2 x 10 Gb
- 1 x 100 Gb

**ESnet**
- Software Defined Networking

**Power Efficiency**
- Ethernet & IB Fabric

**Research Systems Today**
<table>
<thead>
<tr>
<th></th>
<th>Edison</th>
<th>Mira</th>
<th>Titan</th>
<th>Hopper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak Flops (PF)</td>
<td>2.4</td>
<td>10.0</td>
<td>5.26 (CPU) 21.8 (GPU)</td>
<td>1.29</td>
</tr>
<tr>
<td>CPU cores</td>
<td>124,800</td>
<td>786,432</td>
<td>299,008 (CPU) 18,688 (GPU’s)</td>
<td>152,408</td>
</tr>
<tr>
<td>Frequency (GHz)</td>
<td>2.4</td>
<td>1.6</td>
<td>2.2 (CPU) 0.7 (GPU)</td>
<td>2.1</td>
</tr>
<tr>
<td>Memory (TB)</td>
<td>333</td>
<td>786</td>
<td>598 (CPU) 112 (GPU)</td>
<td>217</td>
</tr>
<tr>
<td>Memory/node (GB)</td>
<td>64</td>
<td>16</td>
<td>32 (CPU) 6 (GPU)</td>
<td>32</td>
</tr>
<tr>
<td>Memory BW* (TB/s)</td>
<td>530.4</td>
<td>1406</td>
<td>614 (CPU) 3,270 (GPU)</td>
<td>331</td>
</tr>
<tr>
<td>Memory BW/node* (GB/s)</td>
<td>98</td>
<td>29</td>
<td>33 (CPU) 175 (GPU)</td>
<td>52</td>
</tr>
<tr>
<td>Filesystem</td>
<td>7.6 PB 163 GB/s</td>
<td>35 PB 240 GB/s</td>
<td>10 PB 240 GB/s</td>
<td>2 PB 70 GB/s</td>
</tr>
<tr>
<td>Peak Bisection BW (TB/s)</td>
<td>11.0</td>
<td>24.6</td>
<td>11.2</td>
<td>5.1</td>
</tr>
<tr>
<td>Peak Bisection BW/node (GB/s)</td>
<td>2.12</td>
<td>0.50</td>
<td>0.60</td>
<td>0.80</td>
</tr>
<tr>
<td>Sf ft</td>
<td>1200</td>
<td>~1500</td>
<td>4352</td>
<td>1956</td>
</tr>
</tbody>
</table>
| Power (MW Linpack)       | 2.10   | 3.95  | -12-                         | 8.21   | 2.91
The Computational Research and Theory (CRT) building will be the home for NERSC-8

- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 42 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 Fall 2014**
NERSC-8 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.
Although architecture for NERSC-8 is not yet known, trend is toward manycore processors

- Regardless of chip vendor chosen for NERSC-8, users will need to modify applications to achieve performance
- Multiple levels of code modification may be necessary
  - Expose more on-node parallelism in applications
  - Increase application vectorization capabilities
  - For co-processor architectures, locality directives must be added
# NERSC Upgrades: Meeting Demand

<table>
<thead>
<tr>
<th>System attributes</th>
<th>NERSC-6</th>
<th>NERSC-7</th>
<th>NERSC-8 (proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hopper</td>
<td>Edison</td>
<td></td>
</tr>
<tr>
<td>System peak</td>
<td>1.3 PF</td>
<td>2.4PF</td>
<td>20-40PF</td>
</tr>
<tr>
<td>Power</td>
<td>2.9 MW (Peak)</td>
<td>3 MW (Peak)</td>
<td>&lt;5 MW (Peak)</td>
</tr>
<tr>
<td></td>
<td>2.2MW (Typical)</td>
<td>1.6 MW (Typical)</td>
<td></td>
</tr>
<tr>
<td>System memory</td>
<td>0.21 PB</td>
<td>0.33 PB</td>
<td>1-2 PB</td>
</tr>
<tr>
<td>Node performance</td>
<td>202GF</td>
<td>460 GF</td>
<td>2-3.5TF</td>
</tr>
<tr>
<td>Node memory BW</td>
<td>50 GB/s</td>
<td>100 GB/s</td>
<td>100-500 GB/s</td>
</tr>
<tr>
<td>Node concurrency</td>
<td>24 AMD Magnycours cores</td>
<td>24 Intel Ivy Bridge Cores</td>
<td>up to 512</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>6,384 nodes</td>
<td>5,200 nodes</td>
<td>8,000-12,000 nodes</td>
</tr>
<tr>
<td>MPI Node Interconnect BW</td>
<td>~3 GB/s</td>
<td>~9GB/s</td>
<td>Up to 15GB/s</td>
</tr>
</tbody>
</table>
NERSC’s Application Readiness Strategy

We will use a number of approaches to prepare our diverse user community for the N8 architecture

Vendor/NERSC / ACES partnership
Create a tight partnership with selected NERSC-8 integrator and chip vendor.

Early testbeds for users
NERSC will provide early testbed to users. Many NERSC users are sophisticated and can make progress porting applications independently.

Partner with and leverage existing efforts
Learn from SciDAC engagements, OLCF, ALCF, LLNL application readiness efforts. Exchange lessons learned and best practices.

Developers Workshops
Host a series of developer workshops. Important because NERSC supports a large number of 3rd party applications, particularly in areas of materials science and chemistry

Engage with Application teams
Deep dives with application teams representing key science areas and algorithms to create case studies for all NERSC users

Widespread training series and online modules
Host workshops, online training and create easy to follow online documentation and training modules
Application Readiness team is examining KNC (and GPUs)

Some applications are well suited to the Knight’s architecture, while others will need significant changes to achieve good performance.

- BerkeleyGW kernel is example of code that can benefit from manycore architecture.
- Early prototype KNC hardware roughly equals performance of Sandybridge processor.
- Optimizations for KNC improve performance on Sandybridge.

- Despite improvements from adding OpenMP and vectorization, this multigrid solver will need further restructuring to run on optimally on KNC.
Forecasting
Requirements with six program offices

- Reviews with six program offices every three years
- Program managers invite representative set of users (typically represent >50% of usage)
- Identify science goals and representative use cases
- Based on use cases, work with users to estimate requirements
- Re-scale estimates to account for users not at the meeting (based on current usage)
- Aggregate results across the six offices
- Validate against information from in-depth collaborations, NERSC User Group meetings, user surveys

Tends to underestimate need because we are missing future users

http://www.nersc.gov/science/requirements-reviews/final-reports/
Keeping up with user needs will be a challenge
Keeping up with user needs will be a challenge (cont.)
Future archival storage needs

![Graph showing future archival storage needs with projected gaps and requirements.]
Exponentially increasing data traffic

First petabyte day expected in 2020

Jump driven by data intensive applications

Major improvements in TCP auto-tuning

NERSC daily routed WAN traffic since 2002
Cross Bay Data Transfer

**All NERSC Traffic**

**Photosystem II X-Ray Study**
NERSC users import more data than they export!

- Importing more than 1PB/month
- Exporting more than 1PB/month
Data Analysis is Playing a Key Role in Scientific Discovery
Palomar Transient Factory: Discovered over 2000 spectroscopically confirmed supernovae in the last 5 years, including the youngest and closest Type Ia supernova in past 40 years.

67 refereed publications to-date including 2 in Science Magazine and 4 Nature articles. Processing pipeline runs on NERSC’s systems nightly and makes heavy use of the Science Gateway Nodes to share the data among the collaboration.
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

PI: Kam-Biu Luk (LBNL)
The Planck Mission

- A European Space Agency (+NASA) satellite mission to measure the temperature and polarization of the Cosmic Microwave Background.
  - The echo of the Big Bang: primordial photons have seen it all.
  - Fluctuations encode all of fundamental physics & cosmology.
  - Planck results assumed by all Dark Energy experiments.

- Realizing the full scientific potential of Planck requires very significant computing resources
  - Tiny signal (µK - nK) requires huge data volume for sufficient S/N
  - 72 detectors sampling at 30-180Hz for 2.5 years => $10^{12}$ samples.
  - Analysis depends critically on Monte Carlo methods
    - Simulate and analyze $10^4$ realizations of the entire mission!

- One of Physics World’s Top 10 Breakthroughs of 2013
Materials Project

- **Idea:** Much ‘cheaper’ and faster to pre-screen materials using computations than making them in lab. More than 35,000 inorganic materials calculated in 2 years, coupled with online design and search tools.

- **NERSC** is the simulation engine behind Materials Project and [www.materialsproject.org](http://www.materialsproject.org) is a science gateway hosted at NERSC which connects HPC simulation and data to the web.

- **Users Dec 2013:**

- **Companies that use resource:** Toyota, Sony, Bosch, 3M, Honda, Samsung, LG Chem, Dow Chemicals, GE Global Research, Intermolecular, Applied Materials, Energizer, Advanced Materials, General Motors, Corning, DuPont, Nippon Steel, L’Oreal USA, Caterpillar, HP, Unilever, Lockheed Martin, Texas Instruments, Ford, Bose, Sigma-Aldrich, Siemens, Raytheon, Umicore, Seagate, …
Exascale and Big Data Face the same Computing Challenges
Data deluge at experimental facilities and improved networking will accelerate this trend towards data intensive computing.
DOE experimental facilities are also facing extreme data challenges

- The observational dataset for the Large Synoptic Survey Telescope will be ~100 PB
- The Daya Bay project will require simulations which will use over 128 PB of aggregate memory
- By 2017 ATLAS/CMS will have generated 190 PB
- Light Source Data Projections:
  - 2009: 65 TB/yr
  - 2011: 312 TB/yr
  - 2013: 1.9 PB/yr
  - EB in 2021?
  - NGLS is expected to generate data at a terabit per second

![Graph showing data rate production](image)

Institution | ALS | NSLS | SLAC |
--- | --- | --- | --- |
<2 years | 5.00E+03 | 1.00E+02 | 4.00E+03 |
2-5 years | 1.00E+04 | 5.00E+04 | 1.00E+05 |
> 5 years | 2.00E+06 | 1.00E+05 | 1.00E+06 |

![Graph showing cost per genome](image)
NERSC Strategy
Strategic Objectives

• Meet the ever-growing computing and data needs of our users by
  – providing usable exascale computing and storage systems
  – transitioning SC codes to execute effectively on manycore architectures
  – influencing the computer industry to ensure that future systems meet the mission needs of SC

• Increase the productivity, usability, and impact of DOE’s user facilities by providing comprehensive data systems and services to store, analyze, manage, and share data from those facilities
Unique data-centric resources will be needed

**Compute Intensive Arch**

- Compute
- On-Package DRAM
- Capacity Memory
- On-node-Storage
- In-Rack Storage
- Interconnect
- Global Shared Disk
- Off-System Network

**Goal:** *Maximum computational density and local bandwidth for given power/cost constraint.*

Maximizes bandwidth density near compute

**Data Intensive Arch**

- Goal: *Maximum data capacity and global bandwidth for given power/cost constraint.*

Bring more storage capacity near compute (or conversely embed more compute into the storage).

*Requires software and programming environment support for such a paradigm shift*
NERSC System Plan
Major Technology Changes That Will Improve Usability

• 2015-16 NERSC-8/Trinity
  – High-bandwidth on-package memory
  – “Burst Buffers” – NVRAM enhanced I/O

• 2017-18 CORAL
  – On-die NIC – lower latency
  – On-node NVRAM

• 2019-20 NERSC-9/ATS-3
  – P0 exascale processor
  – Emerging Exascale Programming Model
  – Object-based storage
  – Advanced memory technologies
  – Processing Near Memory (processing data where it is located)
  – Advanced power management technology
  – Coherence domains & fine-grained interprocessor communication

• 2021-22 CORAL+1
  – P1 exascale processor
  – .....
## NERSC Upgrades: Meeting Demand

<table>
<thead>
<tr>
<th>System attributes</th>
<th>NERSC-6</th>
<th>NERSC-7</th>
<th>NERSC-8 (proposed)</th>
<th>NERSC-9 (Proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NERSC-6</strong></td>
<td>Hopper</td>
<td>Edison</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>System peak</strong></td>
<td>1.3 PF</td>
<td>2.6 PF</td>
<td>20-40PF</td>
<td>250-500 PF</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>2.9 MW (Peak)</td>
<td>2.3 MW (Peak)</td>
<td>&lt;5 MW (Peak)</td>
<td>&lt; 15 MW (peak)</td>
</tr>
<tr>
<td></td>
<td>2.2 MW (Typical)</td>
<td>1.6 MW (Typical)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>System memory</strong></td>
<td>0.21 PB</td>
<td>0.35 PB</td>
<td>1-2 PB</td>
<td>~10 PB (128 GB on package, 512-1024 GB DRAM)</td>
</tr>
<tr>
<td><strong>Node performance</strong></td>
<td>202GF</td>
<td>460 GF</td>
<td>2-3.5TF</td>
<td>~10 TF</td>
</tr>
<tr>
<td><strong>Node memory BW</strong></td>
<td>50 GB/s</td>
<td>90 GB/s</td>
<td>100-500 GB/s</td>
<td>~200 GB/s ? 2-4 TB/s on package</td>
</tr>
<tr>
<td><strong>Node concurrency</strong></td>
<td>24 AMD Magnycours cores</td>
<td>24 Intel Ivy Bridge Cores</td>
<td>up to 300</td>
<td>Up to 2048</td>
</tr>
<tr>
<td><strong>System size (nodes)</strong></td>
<td>6,384 nodes</td>
<td>5,576 nodes</td>
<td>8,000-12,000 nodes</td>
<td>O(10,000)</td>
</tr>
<tr>
<td><strong>MPI Node Interconnect BW</strong></td>
<td>~3 GB/s</td>
<td>~9GB/s</td>
<td>~9 GB/s</td>
<td>Up to 50 GB/s</td>
</tr>
</tbody>
</table>