ASCR Facility
Plans

Sudip Dosanjh
Director

June 10, 2015
Projections on Moore’s Law and other trends

- Transistors (Thousands)
- Frequency (MHz)
- Power (W)
- Cores
We need to transition to energy efficient architectures

Manycore or Hybrid is the only approach that crosses the exascale finish line
Projected Parallelism for Exascale

1 billion per cycle
1 million per cycle
1,000 per cycle

Top 10  Top System  Top 1 Trend  Historical  Heavy Node Projections
Can Get Capacity **OR** Bandwidth
But Cannot Get Both in the Same Technology

<table>
<thead>
<tr>
<th>Bandwidth\Capacity</th>
<th>16 GB</th>
<th>32 GB</th>
<th>64 GB</th>
<th>128 GB</th>
<th>256 GB</th>
<th>512 GB</th>
<th>1 TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 TB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 TB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Stack/PNM</td>
</tr>
<tr>
<td>1 TB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Interposer</td>
<td></td>
</tr>
<tr>
<td>512 GB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HMC organic</td>
<td></td>
</tr>
<tr>
<td>256 GB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DIMM</td>
<td></td>
</tr>
<tr>
<td>128 GB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NVRAM</td>
<td></td>
</tr>
</tbody>
</table>

Cost (increases for higher capacity and cost/bit increases with bandwidth)

Power Old Paradigm
• One kind of memory (JEDEC/DDRx)
• ~1 byte per flop memory capacity
• ~1 byte per flop bandwidth

New Paradigm
• DDR4: ~1 byte per flop capacity with <0.01 bytes/flop BW
• Stacked Memory: ~1 byte per flop bandwidth <0.01 bytes/flop capacity
• NVRAM: More capacity, but consumes more Energy for writes than for reads.

1st Step: System Tiering

High Performance Tier

High Capacity Tier

DRAM DRAM DRAM DRAM

HBM

© Samsung
Abstract Machine Model

3D Stacked Memory (Low Capacity, High Bandwidth)

Thin Cores / Accelerators

Fat Core

Core

Coherence Domain

Integrated NIC for Off-Chip Communication

DRAM

NVRAM

(High Capacity, Low Bandwidth)
Leadership Computing for Scientific Discovery

OLCF Titan System Specifications:
- Peak performance of 27.1 Petaflops
  - 24.5 GPU + 2.6 CPU
- 18,688 Hybrid Compute Nodes with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla “K20x” GPU
  - 32 + 6 GB memory
- 200 Cabinets; 710 TB total system memory; 8.9 MW peak power

ALCF Mira System Specifications:
- Peak performance of 10 Petaflops
- 49,152 Compute Nodes each with:
  - 16-Core Power PC A2 CPU with 64 Hardware Threads and 16 Quad FPUs
  - 16 GB memory
- 56 Cabinets; 786 TB total system memory; 4.8 MW peak power

Peer reviewed projects are chosen to advance science, promote innovation, and strengthen industrial competitiveness.

Demand for these machines has grown each year, requiring recent upgrades of both.

FY 2013 research projects include; advancing materials for lithium air batteries, solar cells, and superconductors; improving combustion in fuel-efficient, near-zero-emissions systems; understanding how turbulence affects the efficiency of aircraft and other transportation systems; designing next-generation nuclear reactors and fuels; developing fusion energy systems.
**System Specifications:**

- **Hopper XT5 (2010)**
  - 1.3PF, 212TB, 2.9 MW peak power
- **Edison XC30 (in acc)**
  - Based on DARPA/DOE HPCS system
  - 2.4PF, 333TB, 2.1 MW peak power
- **400TF mixed use clusters**
  - NERSC, JGI, HEP/NP, Materials, Kbase

**Computational Research and Theory Building will provide 12 MW power and cooling for future NERSC computing resources**
## ASCR Computing At a Glance

<table>
<thead>
<tr>
<th>System attributes</th>
<th>NERSC Now</th>
<th>OLCF Now</th>
<th>ALCF Now</th>
<th>NERSC Upgrade</th>
<th>OLCF Upgrade</th>
<th>ALCF Upgrades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name Planned Installation</td>
<td>Edison</td>
<td>TITAN</td>
<td>MIRA</td>
<td>Cori 2016</td>
<td>Summit 2017-2018</td>
<td>Theta</td>
</tr>
<tr>
<td>System peak (PF)</td>
<td>2.6</td>
<td>27</td>
<td>10</td>
<td>&gt; 30</td>
<td>150</td>
<td>&gt;8.5</td>
</tr>
<tr>
<td>Peak Power (MW)</td>
<td>2</td>
<td>9</td>
<td>4.8</td>
<td>&lt; 3.7</td>
<td>10</td>
<td>1.7</td>
</tr>
<tr>
<td>Total system memory</td>
<td>357 TB</td>
<td>710TB</td>
<td>768TB</td>
<td>~1 PB DDR4 + High Bandwidth Memory (HBM) +1.5PB persistent memory</td>
<td>&gt;1.74 PB DDR4 + HBM + 2.8 PB persistent memory</td>
<td>&gt;480 TB DDR4 + High Bandwidth Memory (HBM) Local Memory and Persistent Memory</td>
</tr>
<tr>
<td>Node performance (TF)</td>
<td>0.460</td>
<td>1.452</td>
<td>0.204</td>
<td>&gt; 3</td>
<td>&gt; 40</td>
<td>&gt; 3</td>
</tr>
<tr>
<td>Node processors</td>
<td>Intel Ivy Bridge</td>
<td>AMD Opteron Nvidia Kepler</td>
<td>64-bit PowerPC A2</td>
<td>Intel Knights Landing many core CPUs Intel Haswell CPU in data partition</td>
<td>Multiple IBM Power9 CPUs &amp; multiple Nvidia Voltas GPUS</td>
<td>2nd gen Intel Xeon Phi processor (code name Knights Landing) 3rd gen Intel Xeon Phi processor (code name Knights Hill)</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>5,600 nodes</td>
<td>18,688 nodes</td>
<td>49,152</td>
<td>9,300 nodes 1,900 nodes in data partition</td>
<td>~3,500 nodes</td>
<td>&gt;2,500 nodes</td>
</tr>
<tr>
<td>System Interconnect</td>
<td>Aries</td>
<td>Gemini</td>
<td>5D Torus</td>
<td>Aries</td>
<td>Dual Rail EDR-IB</td>
<td>Aries</td>
</tr>
<tr>
<td>File System</td>
<td>7.6 PB 168 GB/s, Lustre®</td>
<td>32 PB 1 TB/s, Lustre®</td>
<td>26 PB 300 GB/s GPFS™</td>
<td>28 PB 744 GB/s Lustre®</td>
<td>120 PB 1 TB/s GPFS™</td>
<td>10PB, 210 GB/s Lustre initial</td>
</tr>
</tbody>
</table>

### Notes:
- **NERSC Now**
- **OLCF Now**
- **ALCF Now**
- **NERSC Upgrade**
- **OLCF Upgrade**
- **ALCF Upgrades**
- **File System**
- **System attributes**
- **System Interconnect**
- **Node processors**
- **Node performance (TF)**
- **System size (nodes)**
- **Total system memory**
- **Peak Power (MW)**
- **System peak (PF)**

**U.S. Department of Energy**

**Office of Science**

**LANS/LLNS May 29, 2015**
ESnet Goes Global: Extension to Europe

- 25% of all ESnet traffic goes to/from Europe
- 3x100+ Gbps across the Atlantic with redundant paths to serve all DOE missions
- Ready by March 2015 to support LHC Run 2 (was operational in Jan 2015)
- Will support 10x increase in transatlantic traffic from Large Hadron Collider
Recent Developments at NERSC

June 10, 2015
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 2015**
NERSC Timeline

- 2015
  - NRP complete
  - 12.5 MW
  - NERSC-8 Cori Phase I
  - Edison moves

- 2016
  - CRT 20 MW upgrade
  - NERSC-8 Cori Phase II

- 2017
  - CRT 30+ MW upgrade
  - NERSC-9 150-300 Petaflops

- 2020
  - NERSC-10 Capable Exascale for broad Science

- 2024
  - NERSC-11 5-10 Exaflops

- 2028


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
We support a broad user base

• ~6000 users, and we typically add 300-500 per year
• Geographically distributed: 48 states as well as multinational projects
We support a diverse workload

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

• Architected and deployed data platforms including the largest DOE system focused on genomics

• Edison (N7) is the first Cray petascale system with Intel processors, Aries interconnect and Dragonfly topology (serial #1)

• Cori (N8) will be one of the first large Intel KNL systems and will have unique data capabilities
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 9300 Intel Knights Landing compute nodes – mid 2016
  - Self-hosted, (not an accelerator) manycore processor with over 60 cores per node
  - On-package high-bandwidth memory
- Data Intensive Science Support
  - 10 Haswell processor cabinets 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

System named after Gerty Cori, Biochemist and first American woman to receive the Nobel prize in science.
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
Let the hardware automatically manage the integrated on-package memory as an “L3” cache between KNL CPU and external DDR.

Flat Model: Manually manage how your application uses the integrated on-package memory and external DDR for peak performance.

Hybrid Model: Harness the benefits of both cache and flat models by segmenting the integrated on-package memory.

Maximum performance through higher memory bandwidth and flexibility.
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
20 NESAP Tier-1 and Tier-2 codes

**ASCR (2)**
Almgren (LBNL) – **BoxLib AMR** Framework
used in combustion, astrophysics
Trebotich (LBNL) – **Chombo-crunch** for subsurface flow

**BES (5)**
Kent (ORNL) – **Quantum Espresso**
Deslippe (NERSC) – **BerkeleyGW**
Chelikowsky (UT) – **PARSEC** for excited state materials
Bylaska (PNNL) – **NWChem**
Newman (LBNL) – **EMGeo** for geophysical modeling of Earth

**BER (5)**
Smith (ORNL) – **Gromacs**
Molecular Dynamics
Yelick (LBNL) – **Meraculous** genomics
Ringler (LANL) – **MPAS-O**
global ocean modeling
Johansen (LBNL) – **ACME**
global climate
Dennis (NCAR) – **CESM**

**HEP (3)**
Vay (LBNL) – **WARP & Synergia**
accelerator modeling
Toussaint (U Arizona) – **MILC**
Lattice QCD
Habib (ANL) – **HACC** for n-Body cosmology

**NP (3)**
Maris (U. Iowa) – **MFDn**
*ab initio* nuclear structure
Joo (JLAB) – **Chroma**
Lattice QCD
Christ/Karsch (Columbia/BNL) – **DWF/HISQ**
Lattice QCD

**FES (2)**
Jardin (PPPL) – **M3D**
continuum plasma physics
Chang (PPPL) – **XGC1**
PIC plasma
We are launching the NERSC Exascale Science Applications Program (NESAP)

- NESAP components:

  - Strong support from vendors
  - Developer Workshops for 3rd-Party SW
  - Early engagement with code teams
  - Leverage existing community efforts
  - Postdoc Program
  - NERSC training and online modules
  - Early access to N8 technology
DOE Facilities are Facing a Data Deluge

- Astronomy
- Genomics
- Climate
- Physics
- Light Sources
NERSC users import more data than they export!

Importing more than 1PB/month

Exporting more than 1PB/month
Cori Data Enhancements

• Data partition with large memory nodes, software to enable data workflows (including user defined images)
• IO enhancements-- NVRAM nodes on the interconnect fabric for caching, software defined networking
• Larger disk system

Goals are to enable the analysis of large experimental data sets and in-situ analysis coupled to Petascale simulations
NERSC, LANL and Sandia formed a partnership for next-generation supercomputers

• **Alliance for application Performance at Extreme-scale (APEX)**
  – Visible collaboration between ASCR and ASC
  – Strengthen impact on industry
  – Address challenges transitioning applications to advanced manycore architectures with a broader coalition
  – Risk mitigation on technical challenges

• **Successive deployments**
  – Informal partnership on Hopper and Cielo (2010)
  – Cori and Trinity in 2015-2016
  – NERSC-9 and Crossroads in 2020