## Edison Addresses NERSC’s Workload Needs

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>Intel Ivy Bridge 2.6 GHz</td>
<td>Fast, cutting-edge, commodity processor</td>
</tr>
<tr>
<td>Node</td>
<td>Dual-socket, 64 GB 1866 MHz memory</td>
<td>Large memory per node, Excellent memory bandwidth</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Cray Aries, dragonfly topology</td>
<td>Excellent latency &amp; bandwidth, Excellent scaling, Adaptive routing eases congestion</td>
</tr>
<tr>
<td>Storage</td>
<td>6.48 PB, 140 GB/sec I/O bandwidth, 3 file systems</td>
<td>Large, dedicated data storage, High bandwidth; better metadata</td>
</tr>
</tbody>
</table>

*Performance for High Throughput Apps*  
*Performance at Scale*  
*Data & I/O Improvements*
<table>
<thead>
<tr>
<th></th>
<th>NERSC-6 (Hopper)</th>
<th>Edison</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>Cray XE-6</td>
<td>Cray XC30 “Cascade”</td>
</tr>
<tr>
<td>Compute Nodes / Cores</td>
<td>6,384 / 153,216</td>
<td>5,200 / 124,800</td>
</tr>
<tr>
<td>Processor</td>
<td>2 x AMD “Magny Cours” 2.1GHz, 12 core</td>
<td>2 x Intel Ivy Bridge 2.4GHz, 12 core</td>
</tr>
<tr>
<td>Memory</td>
<td>DDR3 1333 MHz</td>
<td>DDR3 1866 MHz</td>
</tr>
<tr>
<td>Memory per Node / Core</td>
<td>32 GB / 1.3 GB</td>
<td>64 GB / 3.2 GB</td>
</tr>
<tr>
<td>Total Memory</td>
<td>217 TB</td>
<td>333 TB</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Gemini (Torus)</td>
<td>Aries (Dragonfly)</td>
</tr>
<tr>
<td>Sustained Performance (SSP)</td>
<td>144 TF</td>
<td>250 TF</td>
</tr>
<tr>
<td>Peak FLOPS</td>
<td>1.28 PF</td>
<td>2.4 PF</td>
</tr>
<tr>
<td>I/O Bandwidth</td>
<td>70 GB/s</td>
<td>&gt;140 GB/s</td>
</tr>
<tr>
<td>I/O Capacity</td>
<td>2 PB</td>
<td>6.48 PB</td>
</tr>
<tr>
<td>File Systems</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Login Nodes</td>
<td>12 x Quad Shanghi/128GB</td>
<td>12 x Quad Sandy Bridge/512GB</td>
</tr>
</tbody>
</table>
System Design

- Primarily water cooled
- One blower assembly for each cabinet pair (group)
- $\leq 75^\circ\text{F}$ water; $\leq 74^\circ\text{F}$ air
- Water coil on right side of each cabinet
Cabinet Design

- 3 chassis / cabinet
- Up to 16 blades/chassis
  - Up to 8 I/O blades
- 4 Nodes/compute blade
  - 2 sockets/node
- 2 single socket nodes/service/IO blade
Edison Compute Node

- Intel Xeon Processor E5-2695 v2
- 2.4 GHz (3.2 GHz max turbo)
- 12 cores / 24 Threads (Hyperthreading)
- Intel AVX extensions
- 22 nm lithography
- 8 Flops / cycle max => 230 Gflops/socket
- 2 sockets per node => 460 Gflops/node
- Intel QPI Speed 16 GB/sec x 2
- 64 GB 1866 MHz memory/node
- ~100 GB/sec memory bandwidth
Edison Node Layout

- Global links, optical (5 links x 2 ports) “Rank-3” 10 Gbps signaling
- Intra-group, backplane (15 links x 1 port) “Rank 1” 12.5 Gbps signaling
- Dual QPI SMP Links
- 4 Channels DDR3
- Intra-group, copper cables (5 links x 3 ports) “Rank-2” 12.5 Gbps signaling
- PCIe-3 16 bits at 8.0 GT/s (16 GB/s) per direction
- Aries 48-port Router 8 ports to 4 NICs, 40 external network ports
Cray XC30 Compute Blade

**SPECIFICATIONS**
- **Module power:** 2014 Watts
- **PDC max. power:** 900 Watt
- **Air flow req.:** 275 cfm
- **Size:** 2.125 in x 12.95 in x 33.5 in
- **Weight:** <40 lbm

[Diagram showing various components such as Faceplate assembly, QPDC 0 PCIe Gen3 Connections, QPDC 1 PCIe Gen3 Connections, Tolopia, 52V-12V Bus Converter (2), Aries, FPGA, Power and Backplane Connector, Inter-group Cable Connectors, Card Edge Stiffener (2)]
Aries Network

• “Dragonfly” topology

• Configured in 3 ranks:
  – Rank-1 is chassis level
  – Rank-2 is cabinet level
  – Rank-3 is system level

• Global bandwidth tuned by number of optical cables (Rank 3)
  – Edison: 11 TB/sec global bandwidth

• Within a 2 cabinet group
  – Minimal routing – 2 hops
  – Non-minimal routing – 4 hops
Aries Rank-1 Network
Aries Rank-2 Network

- 6 backplanes connected with copper wires in a 2-cabinet group: “Rank-2”
- Pairs of Aries connect to an optical fiber to interconnect groups “Rank-3”
- 4 nodes connect to a single Aries
- 16 Aries connected in backplane “Rank-1”

Cabinet
Cascade Network Overview – Rank-3 Network

- An all-to-all pattern is wired between the groups using optical cables (blue network)
- Up to 240 ports are available per 2-cabinet group
- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections

Edison has 546 optical cables in 6-cable bunches at Rank 3.
Application and Development Environment

• Eases adoption by existing users and projects
  – Easy to port and run production codes
• Supports production software applications, libraries, and tools needed by the entire NERSC workload
  – A robust set of programming languages, models
  – A rich set of highly optimized libraries, tools and applications
  – Community and pre-packaged applications
  – Shared-object libraries and socket communication
• Enables effective application performance at scale, single node (high-throughput computing), and everything in between
# Programming Languages and Compilers supported

<table>
<thead>
<tr>
<th>Programming languages</th>
<th>Supported compilers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortran</td>
<td>Cray Compiling Environment (CCE)</td>
</tr>
<tr>
<td>Python, Perl, Shells</td>
<td>Cray Compiling Environment (CCE)</td>
</tr>
<tr>
<td>C</td>
<td>Intel</td>
</tr>
<tr>
<td>Java</td>
<td>GNU</td>
</tr>
<tr>
<td>C++</td>
<td></td>
</tr>
<tr>
<td>Chapel</td>
<td></td>
</tr>
<tr>
<td>UPC</td>
<td></td>
</tr>
</tbody>
</table>

**Default compiler:**
Intel
Supported Programming Models

- MPI
- Cray SHMEM
- OpenMP
- POSIX Threads
- POSIX Shared Memory
- UPC
- Coarray Fortran
- Chapel
Cray Scientific and Math Libraries

Third party scientific libraries
- MUMPS
- SuperLU
- ParMETIS
- HYPRE
- Scotch

LIBSCI
- LAPACK
- ScaLAPACK
- BLACS
- PBLAS

Trilinos

PETSc

IO libraries
- HDF5
- NetCDF
- Parallel-netcdf

HDF5
NetCDF
Parallel-netcdf

MPI-IO Library

Intel MKL

FFTW

DMAPP API for Aries
Development and Performance Tools

Scalable Debuggers
- DDT
- Totalview

Profiling tools
- CrayPat
- Apprentice2
- IPM

Abnormal Termination Processing (ATP)

PAPI
Key Features of Storage

• 3 Lustre v2.2 scratch file systems for spreading user bandwidth needs
  – Spread users among the 2 file systems to evenly distribute load
  – One file reserved for runs with extreme bandwidth needs (up to 70 GB/s to a single file system)

• 2 x the metadata rates from Hopper in aggregate
  – Also isolates metadata performance to 1 of 3 file systems
Energy Efficiency

Energy-efficient design and components allows chiller-free cooling 100% of the year.

First DOE PF system to use year-round chiller-free cooling.

The Bay Area climate allows NERSC to used evaporative cooling for Edison.
Power for Cooling with and without chillers

Annual savings of ~$200K

<table>
<thead>
<tr>
<th></th>
<th>Option 1</th>
<th>Option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooling Tower Fans</td>
<td>19.32</td>
<td>19.32</td>
</tr>
<tr>
<td>CW Pumps</td>
<td>17.38</td>
<td>25.71</td>
</tr>
<tr>
<td>Chiller</td>
<td>0.00</td>
<td>204.98</td>
</tr>
<tr>
<td>TCHW Pumps</td>
<td>17.17</td>
<td>0.00</td>
</tr>
<tr>
<td>CHW Pumps</td>
<td>0.00</td>
<td>19.97</td>
</tr>
</tbody>
</table>
External Nodes

- **esLogin**
  - Quad processor Sandy Bridge
  - 512 GB DDR3 memory
  - 2 dual-port 10GB ethernet
  - 2 dual-port FDR IB HBAs

- **esMS**
  - Management workstation for esLogin nodes
  - Runs Bright Cluster Management software