Introduction to NERSC

An overview of systems, the center, and our way of doing business

January 2012
• National Energy Research Scientific Computing Center
  – Established 1974, first unclassified supercomputer center
  – Original mission: to enable computational science as a complement to magnetically controlled plasma experiment
  – Today’s mission: accelerate scientific discovery by providing production HPC, data, and communications services for research sponsored by the six DOE Office of Science offices.
  – ~4,000 users, ~500 projects; Hundreds of users each day
<table>
<thead>
<tr>
<th>System</th>
<th>Hopper</th>
<th>Franklin</th>
<th>Carver</th>
<th>Euclid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>Compute</td>
<td>Compute</td>
<td>Compute</td>
<td>Analysis</td>
</tr>
<tr>
<td>Nodes</td>
<td>6,384</td>
<td>9,572</td>
<td>1,202</td>
<td>One</td>
</tr>
<tr>
<td>Node Contents</td>
<td>2 CPUs X 12 cores</td>
<td>1 X 4</td>
<td>1,120 @ 2 X 4</td>
<td>8 X 6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>@ 2 X 6</td>
<td></td>
</tr>
<tr>
<td>Total Cores</td>
<td>153,216</td>
<td>38,288</td>
<td>9,920</td>
<td>48</td>
</tr>
<tr>
<td>CPU</td>
<td>AMD Opteron MagnyCours</td>
<td>AMD Opteron Budapest</td>
<td>Intel Nehalem/Westmere</td>
<td>AMD Opteron</td>
</tr>
<tr>
<td>Memory</td>
<td>**</td>
<td>2 GB/core</td>
<td>**</td>
<td>512 GB Total</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Cray “Gemini”</td>
<td>Cray “SeaStar 2+”</td>
<td>4X QDR Infiniband</td>
<td>N/A</td>
</tr>
<tr>
<td>Storage ***</td>
<td>2 PB Lustre</td>
<td>0.4 PB Lustre</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Likely to be retired soon, possibly as soon as late March 2012

Time to migrate to Hopper!
- Beware of decreased memory per core
- Beware of node architecture difference
- Per-core performance approx. the same
- Start thinking about mixed MPI + OpenMP
Other NERSC Systems

- 50-node “Dirac” GPU test bed
- Data transfer nodes dtn01 and dtn02:
  - Optimize WAN transfer between DOE facilities.
  - Reduce load on computational systems’ login and service nodes
- PDSF
• 32 GB DDR3 1333-MHz memory per node, 1.33 GB per core (6,000 nodes)

• 64 GB DDR3 1333-MHz memory per node, 2.66 GB per core (384 nodes)

• **Common Hopper error message:** "OOM killer terminated this process."
  – Your code has attempted to use too much memory.
## Carver Memory

<table>
<thead>
<tr>
<th>Type of Node</th>
<th>Number</th>
<th>Cores / Node</th>
<th>Mem / Node</th>
<th>Mem / Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nehalem 2.67GHz &quot;smallmem&quot;</td>
<td>960</td>
<td>8</td>
<td>24 GB 1333 MHz</td>
<td>3 GB</td>
</tr>
<tr>
<td>Nehalem 2.67GHz &quot;bigmem&quot;</td>
<td>160</td>
<td>8</td>
<td>48 GB 1066 MHz</td>
<td>6 GB</td>
</tr>
<tr>
<td>Westmere 2.67GHz</td>
<td>80</td>
<td>12</td>
<td>48 GB 1333 MHz</td>
<td>4 GB</td>
</tr>
<tr>
<td>Nehalem-EX 2.00GHz</td>
<td>2</td>
<td>32</td>
<td>1 TB 1066 MHz</td>
<td>32 GB</td>
</tr>
</tbody>
</table>
• David and Richard will tell you how to submit jobs so you can target specific memory configurations.
# Hardware Comparisons

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Clock (GHz)</th>
<th>Cores / Node</th>
<th>Peak GFLOPS / s / node</th>
<th>STREAM GB/s/core</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PGI</td>
</tr>
<tr>
<td>Nehalem</td>
<td>2.6</td>
<td>8</td>
<td>83</td>
<td>4391</td>
</tr>
<tr>
<td>Westmere</td>
<td>2.6</td>
<td>12</td>
<td>125</td>
<td>3298</td>
</tr>
<tr>
<td>Magny-Cours (Hopper)</td>
<td>2.1</td>
<td>24</td>
<td>202</td>
<td>2245</td>
</tr>
<tr>
<td>Budapest (Franklin)</td>
<td>2.3</td>
<td>4</td>
<td>37</td>
<td>2298</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hardware</th>
<th>MPI Latency (usec)</th>
<th>MPI Asymptotic Bandwidth (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hopper</td>
<td>1.3 – 2.6</td>
<td>4500</td>
</tr>
<tr>
<td>Carver</td>
<td>1.6</td>
<td>3400</td>
</tr>
<tr>
<td>Franklin</td>
<td>6.2 – 8.4</td>
<td>1700</td>
</tr>
</tbody>
</table>

Caution on performance comparisons - 3 different processor generations
• **Non-Uniform Memory Access**
  - Access to local memory is faster
  - Access to non-local memory is transparent but slower
  - Mostly important for sparsely-packed jobs and MPI / OpenMP
  - Be careful with task placement and memory affinity options (discussed later)

• A single given compute node is always allocated to run a single user job; multiple jobs never share a compute node.
We are working on the exact scope for NERSC-7.
Online Storage Systems

- "Local" file systems
  - Only one system can access
  - "Usually" highest performance

- Global file systems
• Currently Hopper and Franklin only

• Two local file systems on both machines: $SCRATCH$ and $SCRATCH2$

• Lustre file system: designed for high-performance, highly-parallel I/O
  – File per process, MPI-IO, high-level libs, striping considerations

• Franklin 208 TB X 2; Hopper 1 PB X 2

• User quota (0.75 & 5 TB) but increases can be requested

• Not archived! Purged weekly** (all files > 12-weeks access)!

** Purged on Franklin now, starting on Hopper very soon
Center-wide File Systems

- All based on NGF, the NERSC Global Filesystem
- Uses IBM GPFS product
- Architected and managed by NERSC’s Storage Systems Group
- Designed to minimize movement, reduce duplication

- /global/homes

- /global/scratch

- /project

- Also provides /usr/common/
  /usr/common -> /global/common/<platform>
NGF Global Homes

- `/global/homes`: provides common login environment across systems.
  - 50TB total capacity, 15% monthly growth; Tuned for small file access
  - Not purged but archived, quota enforced (40 GB per user), backed up daily
  - Reference it as $HOME; use for source code, small files to save “permanently”
  - Your $HOME directory is shared across all NERSC systems.
NGF Global Scratch

• `/global/scratch`: high bandwidth / capacity TEMPORARY storage
  – Quota enforced (20 TB per user, exceptions granted), **not backed up**!
  – Purged weekly, all files not accessed in 12+weeks!
  – Serves 4000 users, 1PB+ total capacity
  – All users have this automatically; Only scratch system available on Carver and Euclid
  – Tuned for I/O intensive batch jobs, data analysis, viz.; 12GB/s aggregate bandwidth
  – Reference as `$GSCRATCH`
• /project: NERSC-wide sharing and long-term data storage
• Obtain via special request for sharing data between platforms, users, or outside
• Not purged, quota enforced (4TB default per project), backed up daily
• Serves 200 projects; 1.4 PB (+2.8!!) total capacity; ~5 TB average daily IO
• For permanent, archival storage
• Uses magnetic tape, disk with 150TB fast-access disk cache
  – ~15 PB data in 140 M files
  – Increases at ~1.7X per year
  – Average data xfer rate: 100 MB/sec
• Cartridges are loaded unloaded into tape drives by sophisticated robotics
• Use HPSS to back up your code, data
Archival Storage: HPSS

- **HPSS**
  - Access from all NERSC systems + remote
  - Simple unix-like usage via *hsi, htar*
    - pftp, ftp, gridFTP, globus
  - Interactive and / or batch use
  - Help is available for special use cases

* clients available for download
** not ssh
• Compute nodes run applications.

• Service nodes handle support functions.

• Login nodes provide additional user services.
Login nodes should typically be used for the following purposes:
- Develop code (edit, compile/link)
- Submit and monitor batch jobs
- (Some) file management
- Limited interactive post-processing of batch data

- **Carver**: 4 nodes @ 8 cores ea.
- **Hopper**: 12 nodes @ 16 cores ea.
- Login nodes have full OS software environment
• Reached only by use of batch system
  – True for both interactive jobs and jobs without intervention. No direct login access.
  – Use batch system to gain an assignment of compute nodes
• Generally much reduced OS software environment
  – Benefits are better scalability, more user memory
  – OS function availability depends on system: Franklin < Hopper < Carver
• “MOM” nodes
• Reached only by use of batch system
• Used for interactive jobs
  – User launches job
• Also used by the batch system to launch your batch jobs (transparently)
• Reduced OS, especially Franklin, Hopper
• F&H, separate node; C compute node
• Keeping the load down is imperative
Running Jobs

Login Node
Franklin  franklin.nersc.gov
Hopper    hopper.nersc.gov
Carver    carver.nersc.gov

qsub -I -V -q interactive ...

MOM Node
(batch system management)

Franklin: aprun -n ...
Hopper: aprun -n ...
Carver: mpirun -n ...

Compute Node
Compute Node
Compute Node
Compute Node
Compute Node

Franklin: aprun -n ...
Hopper: aprun -n ...
Carver: mpirun -n ...

user

ssh
Service Node Configuration

Hopper

Carver
Choosing a System

- Hopper & Franklin for highly parallel jobs, esp. highly parallel I/O
- Carver memory bandwidth advantage
- OS issues; (No runtime dynamic, shared object libs on Franklin)
- Other queue structure differences
Important Policies

• No production computing using debug / interactive queues.
• No production computing on login nodes.
• No production computing on batch server nodes.
• Do not watch qstat:

```
hopper03 h/hjw> ps | grep watch
1 S root  8340  2  0  80  0 -  0 lcw_di Jan25 ?  00:00:00 [lc_watchdogd]
0 S prxxxxx  22977  16334  0   80  0 -  2463 ?  Jan26 ?  00:02:30 watch qstat -uprxxxxx
0 S hjw  32681  32056  0   80  0 -  1383 pipe_w 17:01 pts/7  00:00:00 grep watch
```
Important Web Page

System Status

http://www.nersc.gov/users/live-status/

LIVE STATUS

<table>
<thead>
<tr>
<th>System</th>
<th>Status</th>
<th>Jobs Running</th>
<th>Cores in Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hooper Compute</td>
<td>UP</td>
<td>238</td>
<td>192,486</td>
</tr>
<tr>
<td>Hooper Login</td>
<td>UP</td>
<td>68</td>
<td>192,386</td>
</tr>
<tr>
<td>Franklin</td>
<td>UP</td>
<td>722</td>
<td>7,600</td>
</tr>
<tr>
<td>Carver</td>
<td>UP</td>
<td>4</td>
<td>208</td>
</tr>
<tr>
<td>PDSF</td>
<td>UP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPSS: archive</td>
<td>UP</td>
<td>(User System)</td>
<td></td>
</tr>
<tr>
<td>HPSS: hpss</td>
<td>UP</td>
<td>(backup)</td>
<td></td>
</tr>
<tr>
<td>Euclid</td>
<td>UP</td>
<td>Interactive</td>
<td></td>
</tr>
</tbody>
</table>

MOTD (Message of the Day)

Contact Information

- NERSC Contacts: http://www.nersc.gov/about/contact-us/
- NERSC Status: http://www.nersc.gov/users/live-status/
- NERSC: 988-66-NERSC (USA) 510-486-8800 (outside continental USA)

Current Status

- Carver: System available.
- Direct: System available.
- Euclid: System available.
- Franklin: System available.
- Hooper: System available.
- HPSS: System available.
- HPSS: hpss: System available.
- Ni: System available.
- NIM: System available.
- PDSF: System available.

Planned Outages

- HPSS User: 02/01/12 09:00-13:00 MT, Scheduled maintenance.
Getting Help

http://www.nersc.gov
1-800-666-3772 (or 1-510-486-8600)
Computer Operations* = menu option 1 (24/7)

<table>
<thead>
<tr>
<th>Account Support = menu option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:accounts@nersc.gov">accounts@nersc.gov</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HPC Consulting = menu option 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:consult@nersc.gov">consult@nersc.gov</a></td>
</tr>
</tbody>
</table>

(8-5, M-F Pacific time)

Online Help Desk = https://help.nersc.gov/

* Passwords during non-business hours
Tips for working with the HPC consultants:

– State which machine your question is about.
– Provide error message(s) if applicable.
– Provide job ID if job crashed
– Provide filesystem, paths to files
– Provide your NERSC user ID
• Make sure you acknowledge NERSC in publications (and talks).

• Science highlights sent to DOE each quarter.
  - Send us links to your publications.
  - See http://www.nersc.gov/news-publications/publications-reports/science-highlights-presentations/
  - See http://www.nersc.gov/news-publications/journal-cover-stories/

1500 publications per year
Thank you.
Additional Info
NERSC at LBNL

- 1000s users, 100s projects
- Allocations:
  - 80% DOE program managers
  - 10% ASCR Leadership Computing Challenge
  - 10% NERSC reserve
- Science includes all of DOE Office of Science
- Machines procured competitively

“Leadership Facilities” at Oak Ridge & Argonne

- 100s users 10s projects
- Allocations:
  - 60% ANL/ORNL managed INCITE process
  - 30% ACSR Leadership Computing Challenge
  - 10% LCF reserve
- Science limited to largest scale; no commitment to DOE/SC offices
- Machines procured through partnerships
## File System Availability

<table>
<thead>
<tr>
<th>System</th>
<th>Hopper</th>
<th>Franklin</th>
<th>Carver</th>
<th>Euclid</th>
<th>PDSF</th>
<th>Datatrans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global home</td>
<td>✔</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Global scratch</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Global Project</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Local Scratch</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## File System Summary

<table>
<thead>
<tr>
<th>File System</th>
<th>Home</th>
<th>Local Scratch</th>
<th>Global Scratch</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope</td>
<td>Global</td>
<td>Local</td>
<td>Global</td>
<td>Global</td>
</tr>
<tr>
<td>Default Quota</td>
<td>40GB 1M inodes</td>
<td>5TB 5M inodes</td>
<td>20TB 2M inodes</td>
<td>4TB 4M inodes</td>
</tr>
<tr>
<td>Intended</td>
<td>• dot files</td>
<td>• batch jobs</td>
<td>• batch jobs</td>
<td>• batch jobs</td>
</tr>
<tr>
<td>Purpose</td>
<td>• source codes</td>
<td>• I/O intensive</td>
<td>• shared access</td>
<td>• shared access</td>
</tr>
<tr>
<td></td>
<td>• compiling</td>
<td>• temporary</td>
<td>• temporary</td>
<td>• temporary</td>
</tr>
<tr>
<td></td>
<td>• input files</td>
<td>storage of large</td>
<td>storage of large</td>
<td>storage of large</td>
</tr>
<tr>
<td>Performance</td>
<td>100MB/sec</td>
<td>35GB/sec</td>
<td>12GB/sec</td>
<td>12GB/sec</td>
</tr>
<tr>
<td>Purged?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Software

- Vendor supplied
- NERSC supplied
- System supplied
- Requests: consult@NERSC.gov
• Cray Data Virtualization Service
• Provides transparent file access to external file systems for processes running on the compute nodes
• At NERSC DVS server nodes connect to NGF and also provide shared-library access
Project
Global Scratch

Hopper with Genimi Network
- CMP
- CMP
- CMP
- CMP
- DVS
- DVS
- DVS

Carver with IB Network
- CMP
- CMP
- CMP
- CMP
- PNSD
- PNSD
- PNSD

Infiniband
- 8
- PNSD
- PNSD
- PNSD

Fiber Channel

5GB/DDN
- DDN 9900
- DDN 9900
- DDN 9900
- DDN 9900

Note: DVS and PNSD are shared between GSCRATCH and PROJECT.

Each DDN 9900 has 300 Disks, Raid6 (8+2) In total 847TB usable disk space
Franklin Scratch

Franklin With SeaStar Network

Fiber Channel

24 OSSs store data on 48 OSTs

Each DDN9500 has 2 controllers, and 160 300GB Disks, Raid6 (8+2)
In total giving 6*160*(8/10)*300GB~209TB

Note: There are two sets of identical configuration for SCRATCH1 and SCRATCH2
• Get involved. Make NUG work for you.
• Provide advice, feedback – we listen.
• Monthly teleconferences with NERSC, usually the last Thursday of the month, 11:00 AM to noon Pacific Time.
• Executive Committee - three representatives from each office and three members-at-large.
• Community!