NERSC Overview

Harvey Wasserman
User Services Group

February 15, 2013
• **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 at the DOE Office of Science through high performance computing and extreme data analysis**
• Diverse workload:
  - 4,500 users, 600 projects
  - 700 codes; 100s of users daily

• Allocations controlled primarily by DOE
  - 80% DOE Annual Production awards (ERCAP):
    - From 10K hour to ~10M hour
    - Proposal-based; DOE chooses
  - 10% DOE ASCR Leadership Computing Challenge
  - 10% NERSC reserve (“NISE”)
DOE View of Workload

NERSC 2013 Allocations By DOE Office

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCR</td>
<td>6%</td>
</tr>
<tr>
<td>BER</td>
<td>18%</td>
</tr>
<tr>
<td>BES</td>
<td>31%</td>
</tr>
<tr>
<td>FES</td>
<td>20%</td>
</tr>
<tr>
<td>HEP</td>
<td>13%</td>
</tr>
<tr>
<td>NP</td>
<td>12%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCR</td>
<td>Advanced Scientific Computing Research</td>
</tr>
<tr>
<td>BER</td>
<td>Biological &amp; Environmental Research</td>
</tr>
<tr>
<td>BES</td>
<td>Basic Energy Sciences</td>
</tr>
<tr>
<td>FES</td>
<td>Fusion Energy Sciences</td>
</tr>
<tr>
<td>HEP</td>
<td>High Energy Physics</td>
</tr>
<tr>
<td>NP</td>
<td>Nuclear Physics</td>
</tr>
</tbody>
</table>
Science View of Workload

NERSC 2013 Allocations
By Science Area
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
Current NERSC Platforms

Large-Scale Computing Systems

Edison (NERSC-7): Cray XC30 (“Cascade”)
- To be operational 2013
- > 2PF/s peak performance; > 0.2 PF/s on apps

Hopper (NERSC-6): Cray XE6
- 6,384 compute nodes, 153,216 cores
- 144 TF/s on applications; 1.3 PF/s peak

Midrange
- 140 Tflops total

Carver
- IBM iDataplex cluster
- 9,884 cores; 106TF peak

PDSF (HEP/NP)
- ~1K core cluster

GenePool (JGI)
- ~5K core cluster
- 2.1 PB Isilon File System

NERSC Global Filesystem (NGF)
- Uses IBM’s GPFS
  - 8.5-PB capacity
  - 15 GB/s global bandwidth

HPSS Archival Storage
- 240 PB capacity
- 5 Tape libraries
- 200TB disk cache

Testbeds

Dirac GPU testbed
(48 nodes)
## Current NERSC Platforms

<table>
<thead>
<tr>
<th>System</th>
<th>Hopper</th>
<th>Carver</th>
<th>Edison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>6,384</td>
<td>1,202</td>
<td>664</td>
</tr>
<tr>
<td>Node Contents</td>
<td>2 CPUs X 12 cores</td>
<td>1,120 @ 2 X 4 @ 2 X 6</td>
<td>2 CPUs X 8 cores</td>
</tr>
<tr>
<td>Total Cores</td>
<td>153,216</td>
<td>9,920</td>
<td>10,624</td>
</tr>
<tr>
<td>CPU</td>
<td>AMD Opteron MC</td>
<td>Intel Nehalem</td>
<td>Intel Sandy Bridge</td>
</tr>
<tr>
<td>Memory</td>
<td>32 GB/node</td>
<td>24 GB/node</td>
<td>64 GB/node</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Cray “Gemini”</td>
<td>4X QDR Infiniband</td>
<td>Cray “Aries” Dragonfly topo</td>
</tr>
<tr>
<td>Storage ***</td>
<td>2 PB Lustre</td>
<td>1.1 PB GPFS</td>
<td>1.6 PB Lustre</td>
</tr>
<tr>
<td>Peak GF/core</td>
<td>8.4</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>Peak GF/node</td>
<td>202</td>
<td>85</td>
<td>332</td>
</tr>
</tbody>
</table>
NERSC Roadmap

Peak Teraflop/s


- Franklin (N5) +QC
  - 36 TF Sustained
  - 352 TF Peak

- Hopper (N6)
  - >1 PF Peak

- Edison (N7)
  - >2 PF Peak

- CRT Facility

- N10 ~1 EF Peak

- N9 200 PF Peak

- N8 75 PF Peak
Chemistry & Materials Applications

• NERSC compiles and supports a large number of software packages for our users.

  ![Number of Scientists using 3rd Party Applications for Simulations](image)

  **A linear-scaling density functional method**

  - More than 13.5 million lines of source code Compiled, Optimized, and Tested

  - Expert advice provided on using these applications
Usage Model

- Compute nodes run applications.
- Service nodes handle support functions.
- Login nodes provide additional user services.
Login Nodes

• **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.
• **Edison:** 6 nodes @ 16 cores ea. + HT
• **Login nodes have full OS software environment**
• **Access:** `ssh -Y -l userId`

![NERSC Logo](https://example.com/nersc_logo.png)

![Office of Science Logo](https://example.com/ofo_science_logo.png)

![NERSC Logo](https://example.com/nersc_logo.png)

![Berkeley Lab Logo](https://example.com/berkeley_lab_logo.png)
• Resource manager and scheduling system assign compute nodes to users
  – Interactive Batch: job is assigned quantity of nodes and shell is started so user can run on those nodes directly; limited time duration (30 min)
  – Non-interactive Batch: job is assigned nodes and batch system then manages the work
    • Consider using an alias for this?

• At NERSC: TORQUE (#PBS) and Moab
• Batch job usage is charged against your NERSC repository
• Access only via batch system
  – True for both interactive jobs and jobs without interactivity.
  – No direct login access.

• Generally reduced OS software environment
  – Benefits are better scalability, more user memory
  – OS function availability depends on system: Hopper < Carver

• Must use job launch command: aprun or mpirun
Running Jobs

- ssh
  - user
  - Login Node
    - Hopper: hopper.nersc.gov
    - Carver: carver.nersc.gov
    - Edison: edison.nersc.gov
  - qsub batch_script.pbs
  - qsub -I -V -q interactive ...

- Compute Nodes
  - Hopper: aprun -n ... a.out
  - Edison: aprun -n ... a.out
  - Carver: mpirun -n ... a.out

- MOM Node
  - (batch system management)
    - Hopper: aprun -n ... a.out
    - Edison: aprun -n ... a.out
    - Carver: mpirun -n ... a.out
Service Nodes

• “MOM” nodes
• Reached only by use of batch system
• Used for interactive jobs
  – User launches job
• Also used by the batch system (transparently) to launch, monitor your batch jobs
• Reduced OS, especially Hopper
• Hopper: separate node; shared with other users
  – Keeping the load down is imperative
• Carver: compute node; yours during the batch job
Batch Scripts

1. Begin `#!/user/bin/csh` or `#!/usr/bin/bash`
2. `#PBS` batch directives
3. Shell commands
4. `aprun` or `mpirun` job launch commands

```bash
#!/usr/bin/csh
#PBS -q regular
#PBS -l mppwidth=16
#PBS -l walltime=00:15:00
#PBS -N SweepEdison
#PBS -j eo

cd $PBS_O_WORKDIR

setenv NO_STOP_MESSAGE zz
aprun -n 16 ./sweep3d.mpi.ed
```
How Usage is Charged

- Elapsed wall-clock time in hours
- Number of nodes allocated to the job (regardless of the number actually used)
- Queue charge factor (QCF).
  - Charge classes: Premium, regular, discounted regular (Hopper only), or low
- Number of cores per allocated node
- Machine charge factor (MCF) based on typical performance of the machine relative to Hopper (MCF=1.0 or 1.5 for Carver)
Special Queues

• **Debugging and interactive jobs**
  – Use –q debug or –q interactive
  – Enabled 5am - 6pm Pacific Time
  – Relative priority higher than regular
  – **Do not use for production work!** (NERSC staff monitors)

• **Hopper reg_xbig queue**
  – Runs Friday 9pm (jobs submitted >48 hours in advance)

• **A variety of other special queues; check**
  [www.nersc.gov](http://www.nersc.gov)
• Two execution modes: native (default) and CCM

• CCM: Cluster Compatibility Mode
  – Used for pre-existing executables;
  – Software that cannot be compiled;
  – Software that needs full Linux software stack, TCP/IP
  – Gaussian, NAMD replica exchange, WIEN2k

• Native: also called Extreme Scalability Mode; 99% of users should use this mode
Simple Rules for Success

• Use compiler wrappers (NOT gcc, g++, ifort, pgf90)

• Submit script or command line to batch system

• Launch job on compute nodes with aprun / mpirun

• Select appropriate file system
NERSC File Systems

• Permanent Data
  – Global “Home” directories: $HOME; not for output from running jobs; backed up; 40 GB limit (no exceptions)
  – HPSS: access via hsi, htar, grid, ftp/pftp
    • Primary backup/archive/permanent storage
  – Project directories: optional; for sharing among users, NERSC systems, external users

• Scratch or temporary data
  – Local scratch: Hopper (5-TB limit), Edison (TBD): Lustre
    • large, high-performance, esp. parallel I/O; purged, not backed up
  – Global scratch: all systems, GPFS (40-TB limit)
NERSC File Systems

- **NERSC Scratch file systems**
  - **Hopper:**
    - Local: `$SCRATCH` and `$SCRATCH2`
    - Global: `$GSCRATCH`
  - **Edison:**
    - Local: `$SCRATCH` (currently)
    - Global: `$GSCRATCH`
  - **Carver:**
    - Global: `$GSCRATCH == $SCRATCH`
NERSC File Systems Suggestions

• Optimal file space is tradeoff between I/O performance, usability
• Use $SCRATCH for highly-parallel, large-scale I/O
• Performance is often a function of metadata:
  – Do not store many (1000s of) files in single directory
• HPSS to archive important data
• Try to optimize I/O if using your own code
  – Large-block reads/writes
  – Advanced options for Lustre (see NERSC web)
  – Use high-level libraries
### NOW COMPUTING

National Energy Research Scientific Computing Center

#### Advanced Simulation of Pore Scale Reactive Transport Processes Associated with Carbon Sequestration (m1516)

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Investigator</th>
<th>Compute Cores</th>
<th>Core Hrs Requested</th>
<th>Core Hours Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computing Research</td>
<td>David Trebotich</td>
<td>49,152</td>
<td>1,769,472</td>
<td>641,532.3</td>
</tr>
<tr>
<td>NSE project</td>
<td>Lawrence Berkeley National Laboratory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geoscience</td>
<td>Computer Hopper</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Parallel Simulation of Electron Cooling Physics and Beam Transport (m327)

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Investigator</th>
<th>Compute Cores</th>
<th>Core Hrs Requested</th>
<th>Core Hours Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear Physics</td>
<td>Kevin Paul</td>
<td>6,144</td>
<td>3,072</td>
<td>2,116.7</td>
</tr>
<tr>
<td>SciDAC</td>
<td>Tech-X Corporation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accelerator Science</td>
<td>Computer Hopper</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Projections of Ice Sheet Evolution Using Advanced Ice and Ocean Models (m1343)

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Investigator</th>
<th>Compute Cores</th>
<th>Core Hrs Requested</th>
<th>Core Hours Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological &amp; Environmental Research</td>
<td>William D. Collins</td>
<td>3,840</td>
<td>30,720</td>
<td>3,498.9</td>
</tr>
<tr>
<td>ALCOL project</td>
<td>Lawrence Berkeley National Laboratory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Climate Research</td>
<td>Computer Hopper</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Central Engine Models for Core-Collapse Supernovae and Long Gamma-Ray Bursts (m152)

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Investigator</th>
<th>Compute Cores</th>
<th>Core Hrs Requested</th>
<th>Core Hours Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear Physics</td>
<td>Christian D. Ott</td>
<td>2,880</td>
<td>138,240</td>
<td>2,529.8</td>
</tr>
<tr>
<td>DOE Base</td>
<td>Louisiana State University</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Astrophysics</td>
<td>Computer Hopper</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Numerical Simulations of Defects and Chemical Reactions at Surfaces and Interfaces (m647)

<table>
<thead>
<tr>
<th>DOE Office</th>
<th>Investigator</th>
<th>Compute Cores</th>
<th>Core Hrs Requested</th>
<th>Core Hours Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic Energy Sciences</td>
<td>Sergey Rashkeev</td>
<td>192</td>
<td>4,608</td>
<td>2,549.2</td>
</tr>
<tr>
<td>DOE Base</td>
<td>Idaho National Laboratory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Materials Science</td>
<td>Computer Carver</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Getting Help

http://www.nersc.gov
http://www.nersc.gov/users/computational-systems/

https://help.nersc.gov

1-800-666-3772 (or 1-510-486-8600)

Account Support = menu option 2
accounts@nersc.gov

HPC Consulting = menu option 3
consult@nersc.gov
(8-5, M-F Pacific Time)

Passwords during non-business hours: call Computer Operations = menu option 1 (24/7)
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; please use “official” acknowledgement

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

1500 publications per year
Thank you.
Additional Info
Logging In

- term% ssh –Y your_login_id@carver.nersc.gov

- Customizations: put in .cshrc.ext or .bashrc.ext

- Do NOT edit .cshrc, .bashrc, .profile, .login
Module Utility

• Most software accessed via module utility

• Why:
  – Allows NERSC to make many installations and versions available
  – Allows user to easily change environment

• What:
  – Consists of module command and modulefiles
Module Utility

- See what’s available:
  - module avail or module avail modulefile_name
  - Example: module avail pgi

- See what you have loaded in your environment:
  - module list

- Load a new one (bring it into your environment):
  - module load modulefile_name

- Swap:
  - module swap old_modulefile_name new_modulefile_name

- See what’s included:
  - module show modulefile_name

- Some basic help:
  - module help modulefile_name
• **Compiling environments: PGI, Intel, GNU**
  
  – Default is PGI on all NERSC systems (except Edison)
  
  – Consist of “native” compilers + all libraries
  
  – Example on Carver:
    
    • Serial compile: `pgf90 my_serial_code.f`
      `ifort my_serial_code.f`
      `gfortran my_serial_code.f`

• **MPI environments: PGI, Intel, GNU**
  
  – Default is PGI on all NERSC systems (except Edison)
  
  – Consist of mpi wrappers, libs, headers

    • MPI compile (all three): `mpif90 my_mpi_code.f`
      `mpicc mpi_code.c`
      `mpiCC/mpiC++/mpicxx mpi_code.C`
• Default is pgi on Hopper, Carver; Carver change requires TWO steps:

- `carver% module swap pgi intel`
- `carver% module swap openmpi openmpi-intel`

- or

- `carver% module swap pgi gcc`
- `carver% module swap openmpi openmpi-gcc`
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
Hopper & Carver Memory

- Use the batch system to submit jobs so you can target specific memory configurations.
Hopper Memory

- 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>

Carver top view
# Hardware Comparisons

<table>
<thead>
<tr>
<th></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>PGI</td>
<td>Intel</td>
<td>Cray</td>
<td>GCC</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></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
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
NGF Project

- /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
Archival Storage: HPSS

- 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
## File System Availability

<table>
<thead>
<tr>
<th>System</th>
<th>Hopper</th>
<th>Edison</th>
<th>Carver</th>
<th>Gene pool</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>

- Not all Edison scratch file systems installed yet
Software

• Vendor supplied
• NERSC supplied
• System supplied
• Requests: consult@NERSC.gov
Hopper Scratch

![Diagram of Hopper Scratch system]

- **Hopper With Gemini Network**
  - CMP nodes interconnected through LNETs

- **Infiniband**
  - OSS nodes connected

- **Fiber Channel**
  - LSI 7900 controllers with 120 1TB disks in Raid6 (8+2) configuration
  - In total giving 13*120*(8/10)*1TB~1.1PB

- **SCRATCH1/2**
  - 26 OSSs store data on 156 OSTs

Note: There are two sets of identical configuration for SCRATCH1 and SCRATCH2.
Hopper with Gemini Network

Franklin with SeaStar Network

Carver/Mag with IB Network

Infiniband

Fiber Channel

5GB/DDN

DVS

PNSD

DVS

PNSD

PNSD

16

20

14

8

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

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

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
NERSC User’s Group

- 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!
• **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.