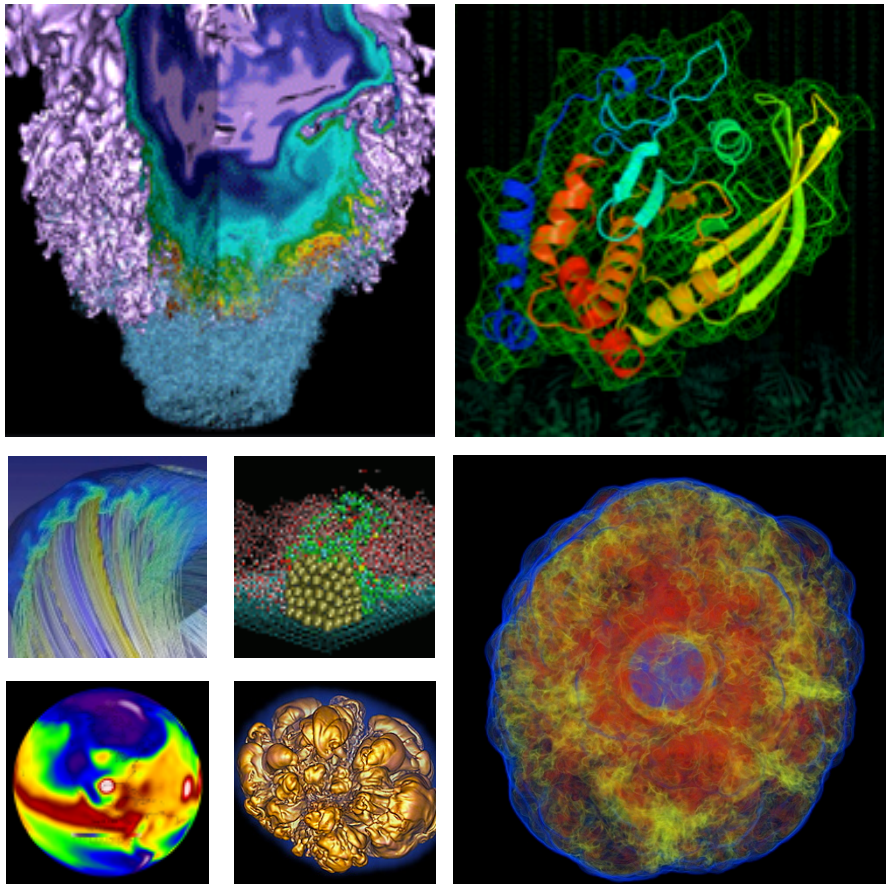
NERSC Overview



Harvey Wasserman
User Services Group

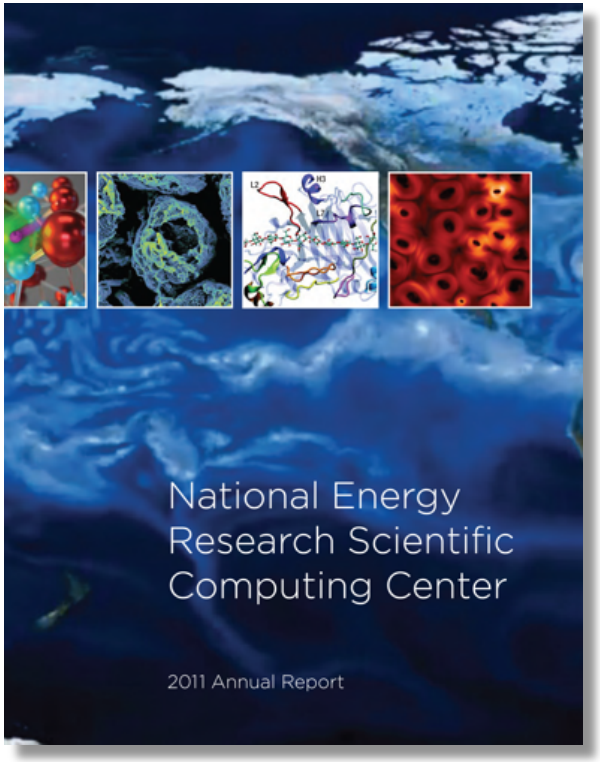
February 15, 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

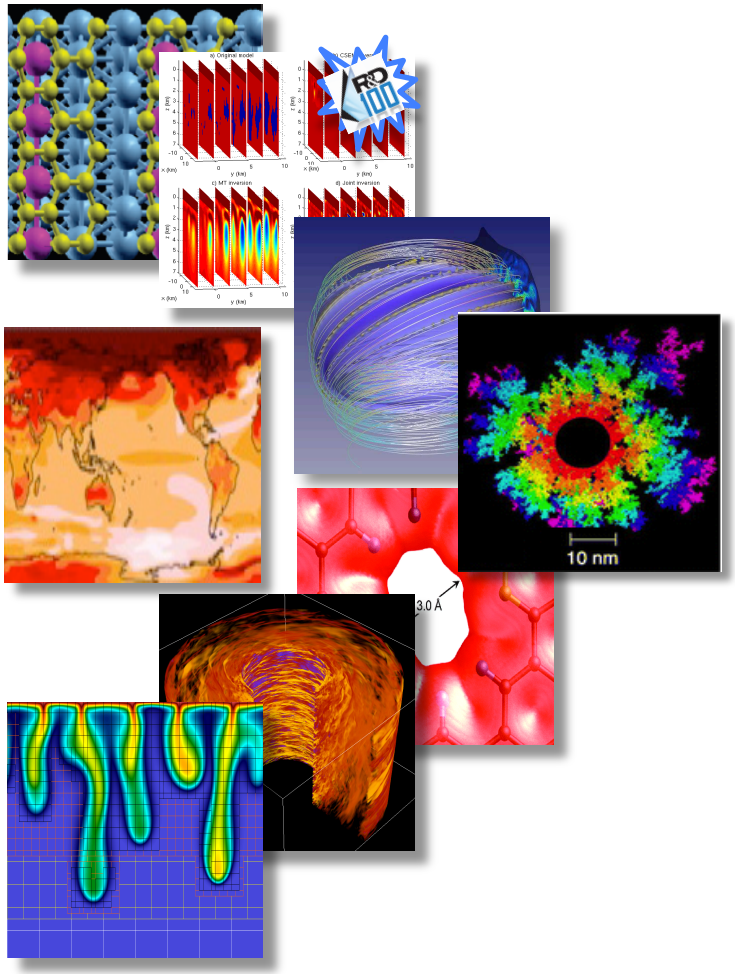
NERSC

- **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**

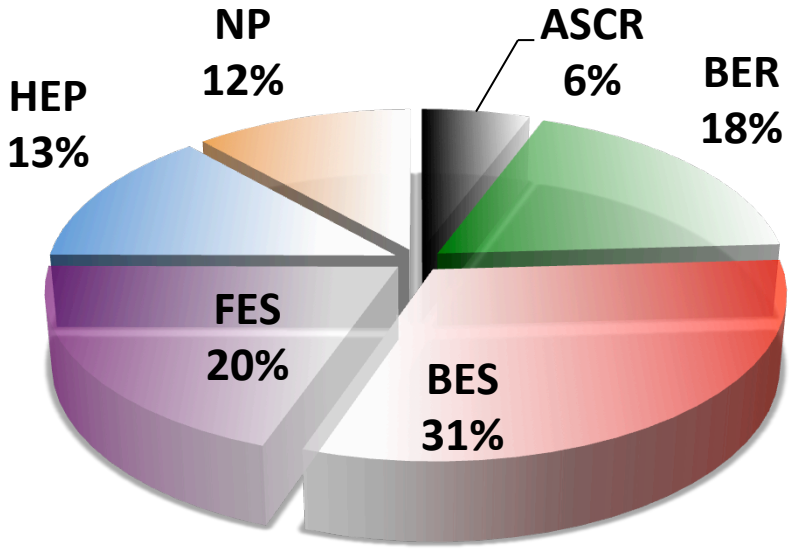


NERSC: Production Computing for the DOE Office of Science

- **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")



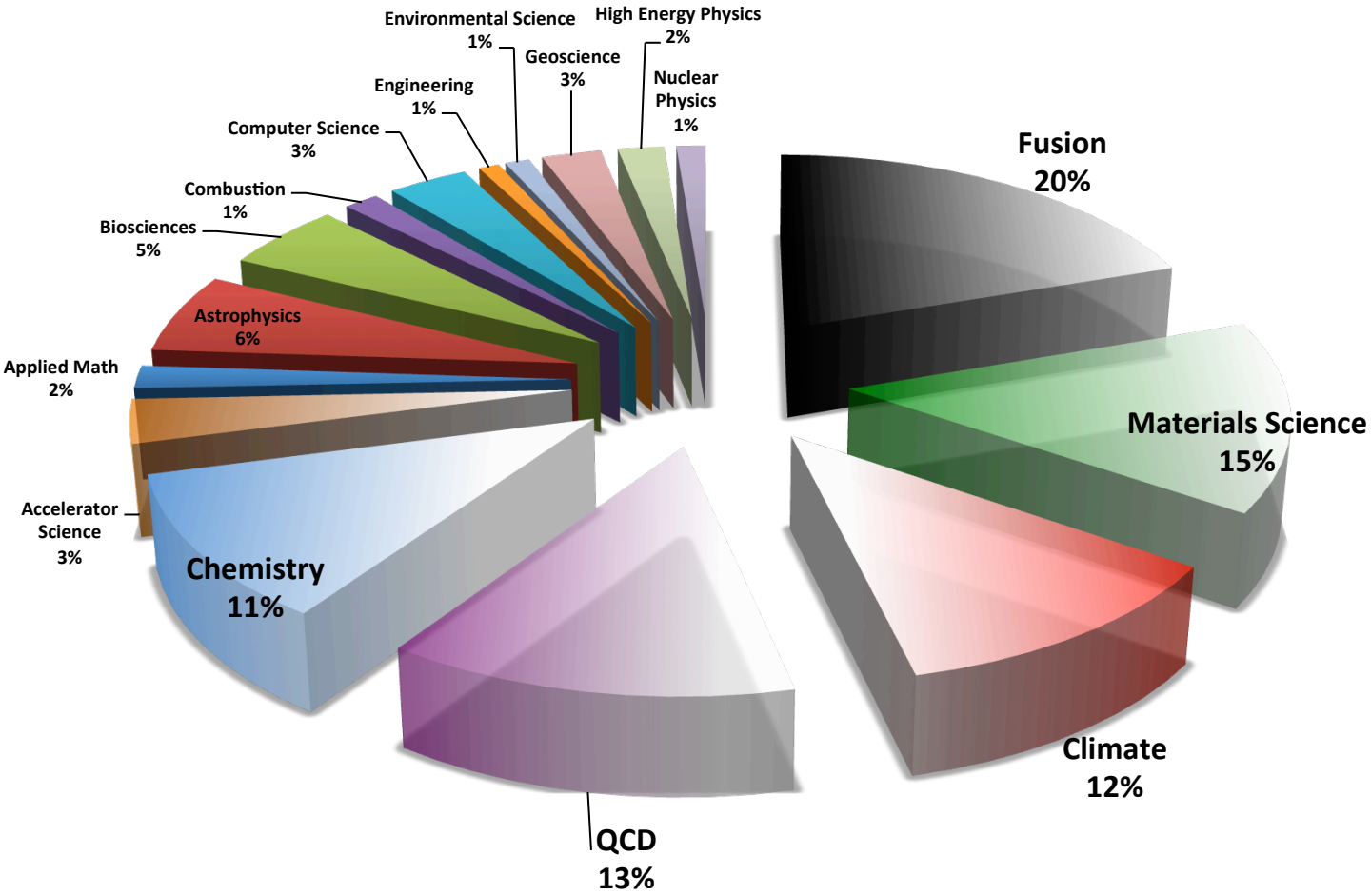
DOE View of Workload



NERSC 2013 Allocations By DOE Office

ASCR	Advanced Scientific Computing Research
BER	Biological & Environmental Research
BES	Basic Energy Sciences
FES	Fusion Energy Sciences
HEP	High Energy Physics
NP	Nuclear Physics

Science View of Workload



NERSC 2013 Allocations By Science Area

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

ASCR Facilities

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

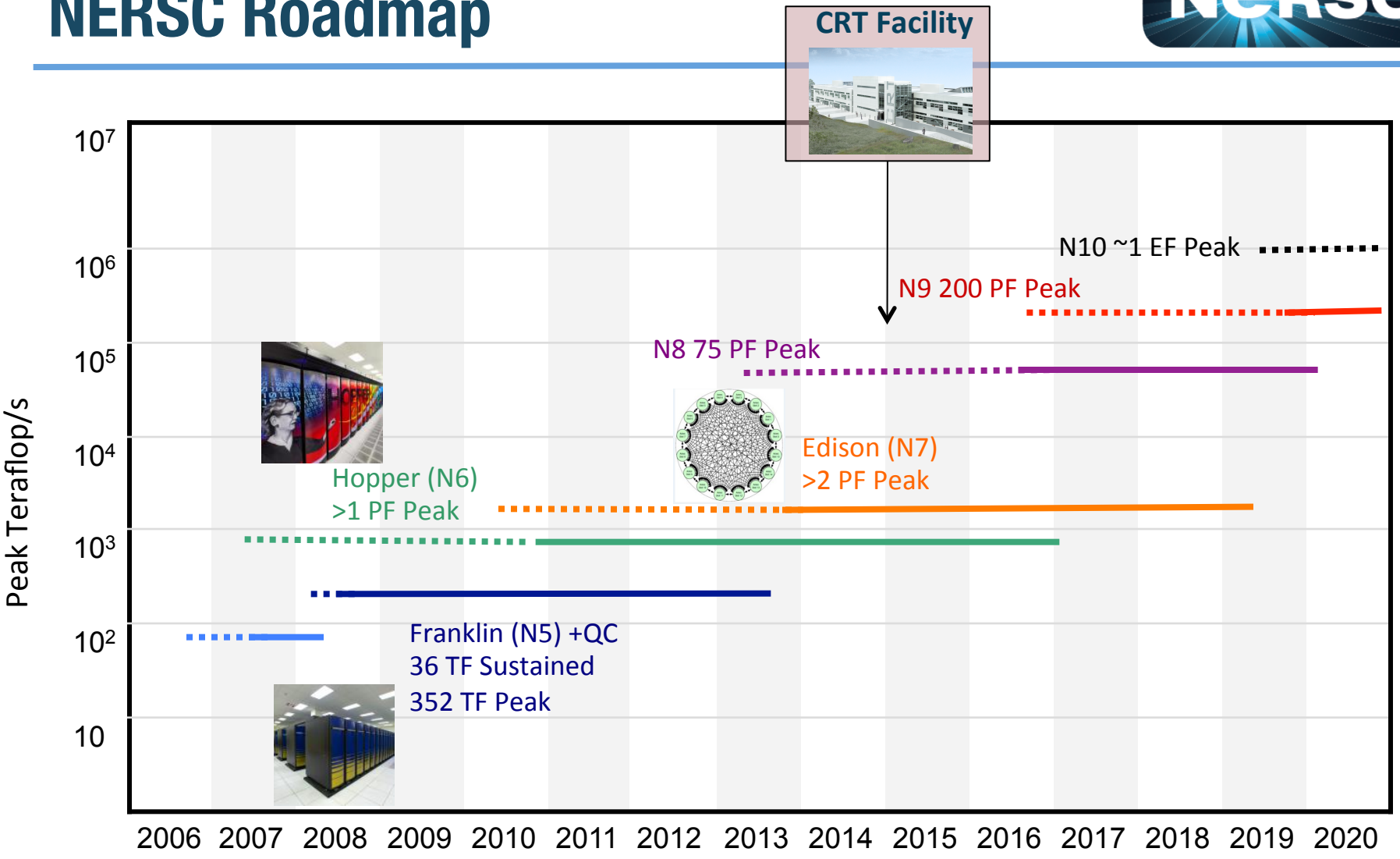
System	Hopper	Carver	Edison
Nodes	6,384	1,202	664
Node Contents	2 CPUs X 12 cores	1,120 @ 2 X 4 80 @ 2 X 6	2 CPUs X 8 cores
Total Cores	153,216	9,920	10,624
CPU	AMD Opteron MC	Intel Nehalem	Intel Sandy Bridge
Memory	32 GB/node	24 GB/node	64 GB/node
Interconnect	Cray "Gemini"	4X QDR Infiniband	Cray "Aries" Dragonfly topo
Storage ***	2 PB Lustre	1.1 PB GPFS	1.6 PB Lustre
Peak GF/core	8.4	11	21
Peak GF/node	202	85	332

Hopper


Carver

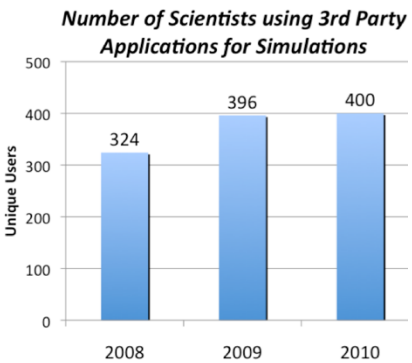
Edison

NERSC Roadmap



Chemistry & Materials Applications

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



A linear-scaling density functional method

Qbox

LAMMPS

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

- **Expert advice provided on using these applications**

NWCHEM

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`**

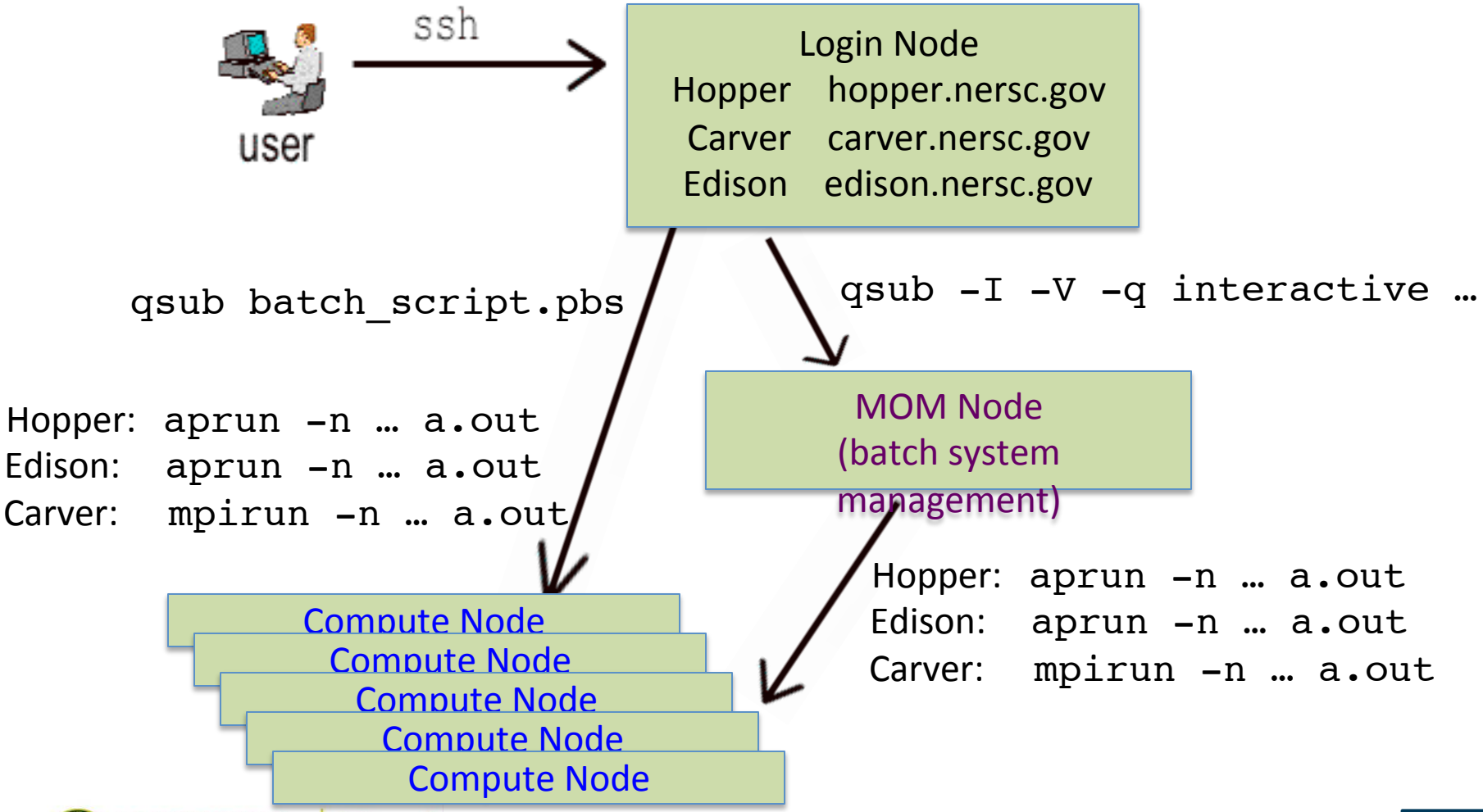
Batch Computing

- **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 respository**

Compute Nodes

- **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



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

```
#!/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**

Running on Cray Systems

- **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 Scalabilility 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

Important Web Page

System Status

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

Google

Login

Site Map | My NERSC | Share

search...

NERSC

Powering Scientific Discovery Since 1974

HOME | ABOUT | SCIENCE AT NERSC | SYSTEMS | FOR USERS | NEWS & PUBLICATIONS | R & D | EVENTS | LIVE STATUS

FOR USERS

- System Status
- Global Queue Look
- Outage Log
- Now Computing Highlights
- My NERSC
- Getting Started
- Help
- Computational Systems
- Queues and Policies
- Job Information
- Software
- Accounts & Allocations
- Analytics & Visualization
- Data & Networking
- Science Gateways
- Training & Tutorials
- User Announcements
- User Surveys
- NERSC Users Group

Passwords & Account Support
https://nim.nersc.gov
accounts@nersc.gov
1-800-66-NERSC, option 2
or 510 486-8612

Consulting
http://help.nersc.gov
consult@nersc.gov
1-800-66-NERSC, option 3
or 510 486-8611

Home » For Users » System Status

LIVE STATUS

Current MOTD

System	Status	Jobs Running	Cores in Use	Description/Notes
Carver:	Up	222	7034	
Dirac:	Up	11	232	
Edison:	Up			
Euclid:	Down			01/31/13 9:00 PST Unavailable. (01/31/13 14:31 PST) - Euclid has been disabled at 09:00 02/01/13. All running jobs and connections are terminated due to system retirement 1/31/13.
Genepool:	Up			
Hopper:	Up	409	149160	
HPSS Backup:	Up			
HPSS User:	Up			
Jesup:	Up			
NGF:	Up			
PDSF:	Up			

Service Status

All services are available.

Planned Outages

Genepool:	02/11/13 0:01-02/18/13 23:59 PST Scheduled maintenance. Genepool phase 2 nodes will be unavailable for one week for planned maintenance.
Shibboleth:	02/11/13 11:00-12:00 PST Scheduled maintenance. shib.nersc.gov, the Login server for some NERSC websites, including www.nersc.gov may be briefly affected.

U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Interesting Web Page

Now Computi

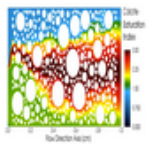
www.nersc.gov/display

NERSC

NOW COMPUTING

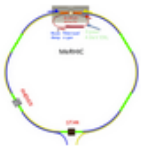
National Energy Research Scientific Computing Center

U.S. DEPARTMENT OF ENERGY | Office of Science



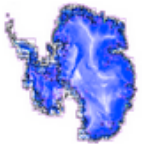
Advanced Simulation of Pore Scale Reactive Transport Processes Associated with Carbon Sequestration (m1516)

DOE Office	Advanced Scientific Computing Research	Investigator	David Trebotich	Compute Cores	49,152
Project Class	NISE project	Organization	Lawrence Berkeley National Laboratory	Core Hrs Requested	1,769,472
Science Area	Geoscience	Computer	Hopper	Core Hours Used	641,532.3



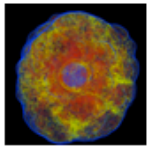
Parallel Simulation of Electron Cooling Physics and Beam Transport (m327)

DOE Office	Nuclear Physics	Investigator	Kevin Paul	Compute Cores	6,144
Project Class	SciDAC	Organization	Tech-X Corporation	Core Hrs Requested	3,072
Science Area	Accelerator Science	Computer	Hopper	Core Hours Used	2,116.7



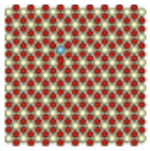
Projections of Ice Sheet Evolution Using Advanced Ice and Ocean Models (m1343)

DOE Office	Biological & Environmental Research	Investigator	William D. Collins	Compute Cores	3,840
Project Class	ALCC project	Organization	Lawrence Berkeley National Laboratory	Core Hrs Requested	30,720
Science Area	Climate Research	Computer	Hopper	Core Hours Used	3,498.9



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

DOE Office	Nuclear Physics	Investigator	Christian D. Ott	Compute Cores	2,880
Project Class	DOE Base	Organization	Louisiana State University	Core Hrs Requested	138,240
Science Area	Astrophysics	Computer	Hopper	Core Hours Used	2,629.8



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

DOE Office	Basic Energy Sciences	Investigator	Sergey Rashkeev	Compute Cores	192
Project Class	DOE Base	Organization	Idaho National Laboratory	Core Hrs Requested	4,608
Science Area	Materials Science	Computer	Carver	Core Hours Used	2,549.2

HPC Systems

Hopper

Cray XE6
Peak TFlop/s: 1,288
Installed: 2011

Cores in Use: 148,368 (96.8 %)
Jobs Running: 446
Jobs Queued: 2,087
Backlog: 25,270,152 core hours

Carver

IBM iDataPlex
Peak TFlop/s: 34
Installed: 2010

Cores in Use: 7,372 (93.0 %)
Jobs Running: 606
Jobs Queued: 378
Backlog: 604,384 core hours

BERKELEY LAB

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)

Getting Help

- **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
  - New issue? New trouble ticket.

Science

- **Make sure you acknowledge NERSC in publications; please use "official" acknowledgement**

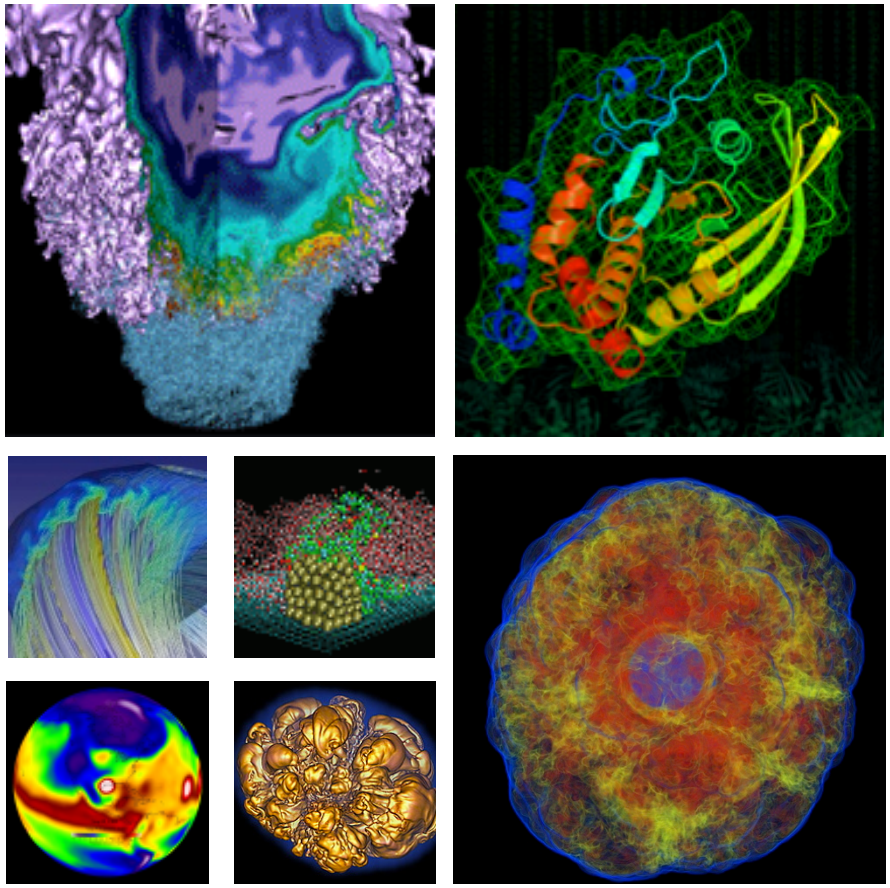
- **Science highlights sent to DOE each quarter.**

  - Send us links to your publications.

  - See http://www.nersc.gov/news-publications/news/
  - 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.

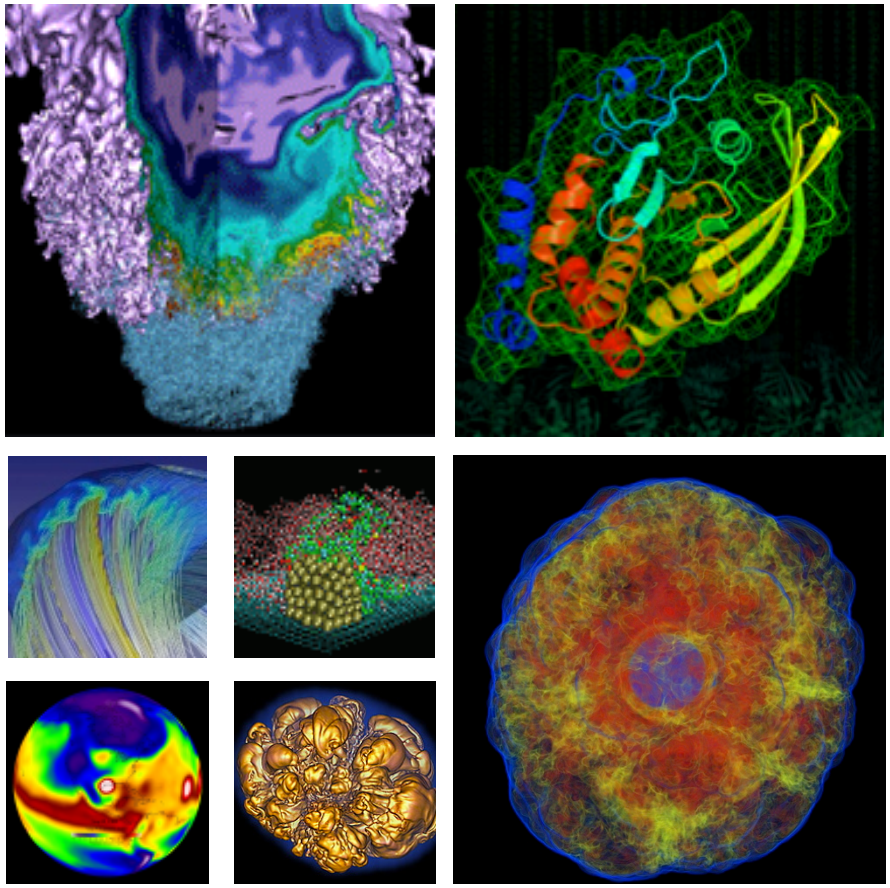


February 15, 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Additional Info



February 15, 2013

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

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

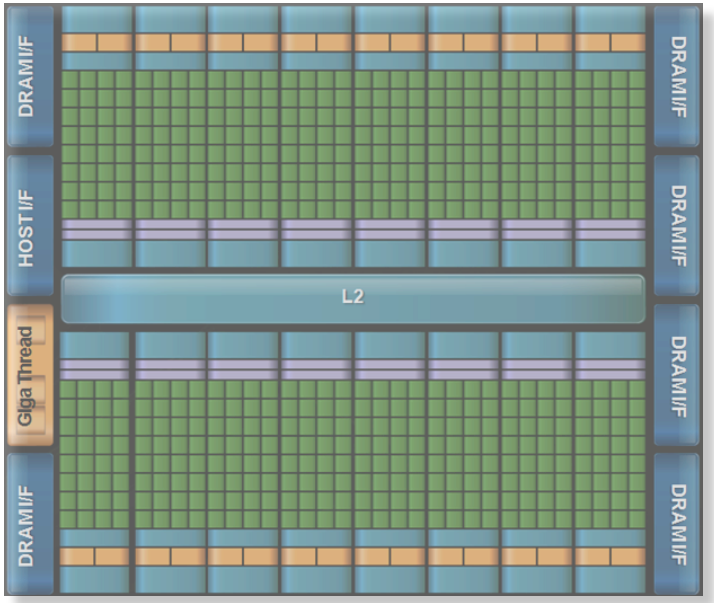
- **Compiling environments: PGI, Intel, GNU**
  - Default is PGI on all NERSC systems (except Edison)
  - Consist of "native" compilers + all libraries
  - Example on Carver:
    - Serial compile: `pgf90 my_serial_code.f`
      `ifort my_serial_code.f`
      `gfortran my_serial_code.f`
- **MPI environments: 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`

Changing Compiler Environments

- **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**

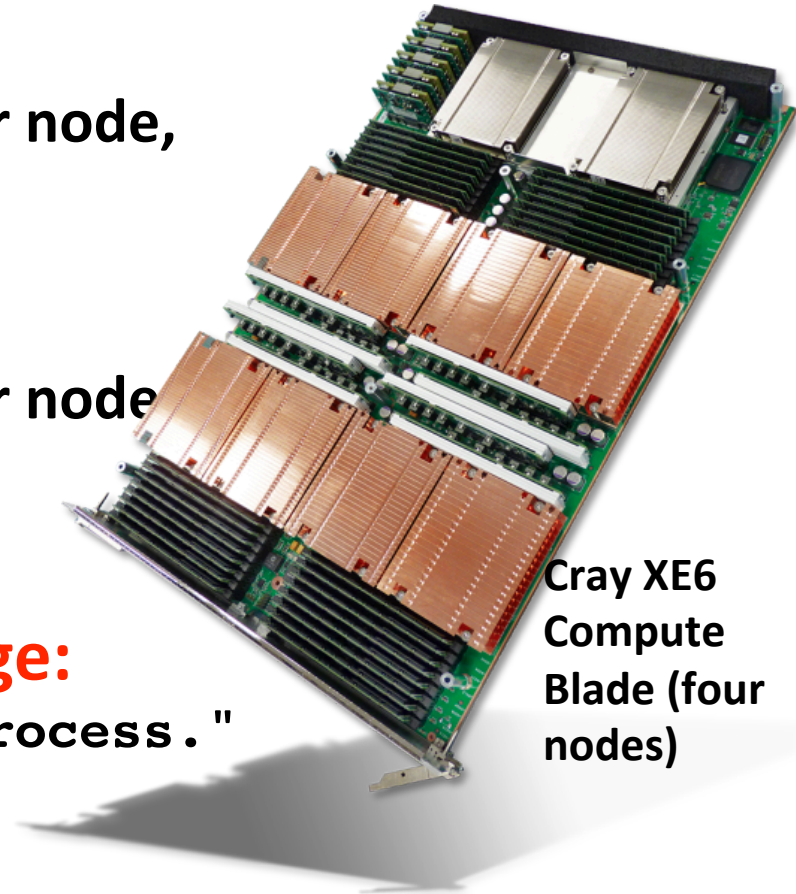


Hopper & Carver Memory

- **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.



Cray XE6 Compute Blade (four nodes)

Carver Memory

Type of Node	Number	Cores / Node	Mem / Node	Mem / Core
Nehalem 2.67GHz "smallmem"	960	8	24 GB 1333 MHz	3 GB
Nehalem 2.67GHz "bigmem"	160	8	48 GB 1066 MHz	6 GB
Westmere 2.67GHz	80	12	48 GB 1333 MHz	4 GB
Nehalem-EX 2.00GHz	2	32	1 TB 1066 MHz	32 GB

Carver top view

Hardware Comparisons

	Clock (GHz)	Cores / Node	Peak GFLOPS / s / node	STREAM GB/s/core			
				PGI	Intel	Cray	GCC
Nehalem	2.6	8	83	4391	4628		
Westmere	2.6	12	125	3298	3516		
Magny-Cours (Hopper)	2.1	24	202	2245	2254	2118	1616
Budapest (Franklin)	2.3	4	37	2298			

	MPI Latency (usec)	MPI Asymptotic Bandwidth (GB/s)
Hopper	1.3 – 2.6	4500
Carver	1.6	3400
Franklin	6.2 – 8.4	1700

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

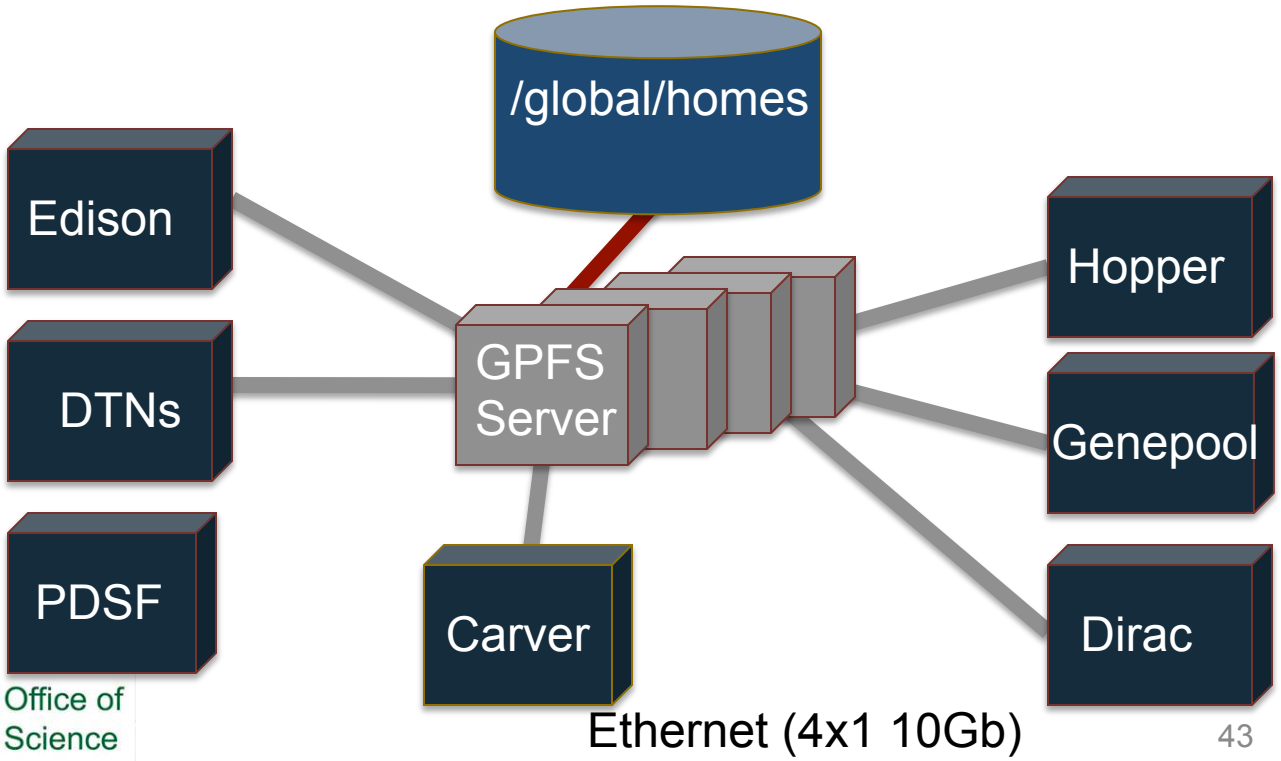
- /global/scratch

- /project

- Also provides `/usr/common/`
  `/usr/common -> /global/common/<platform>`

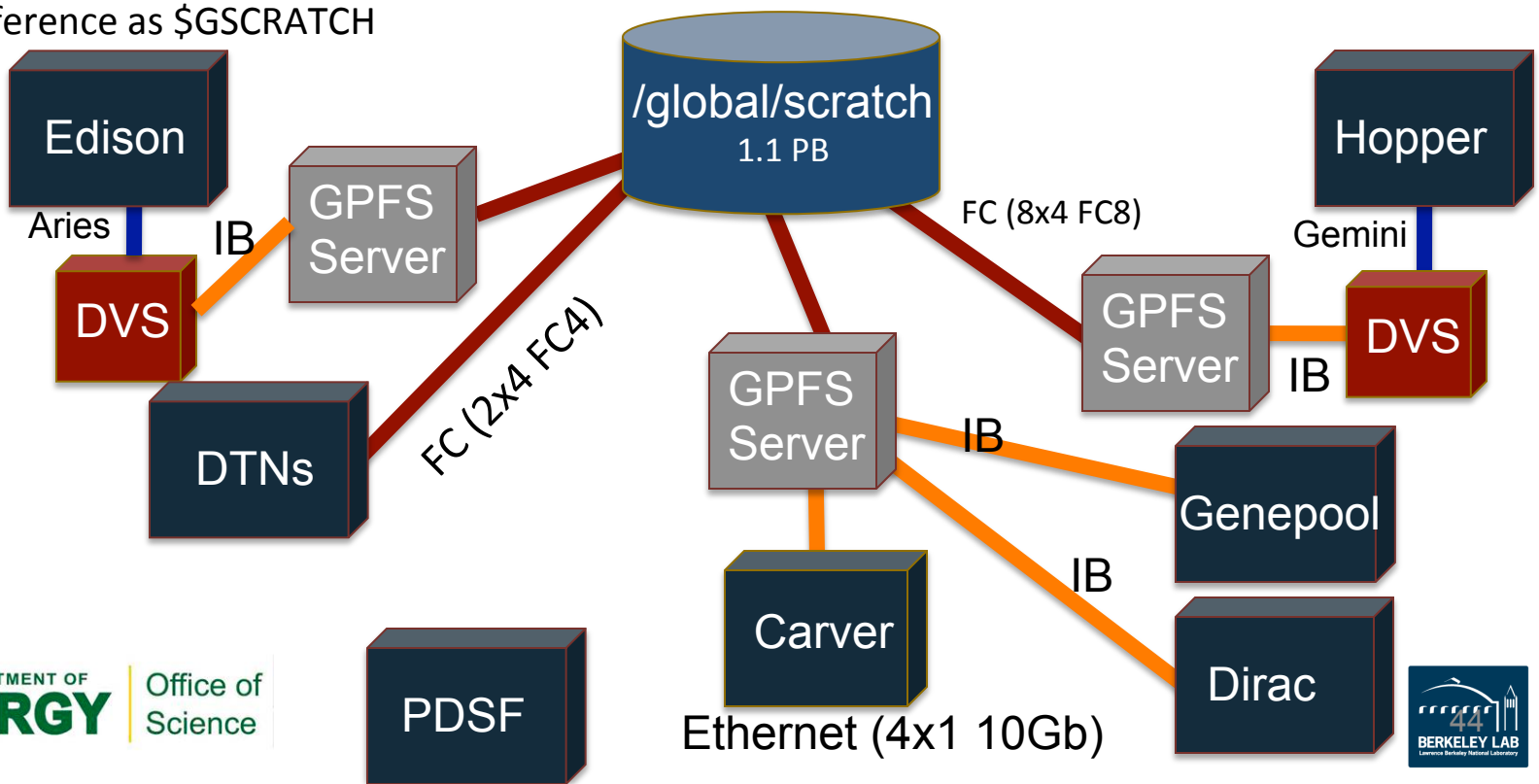
NGF Global Homes

- **/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.



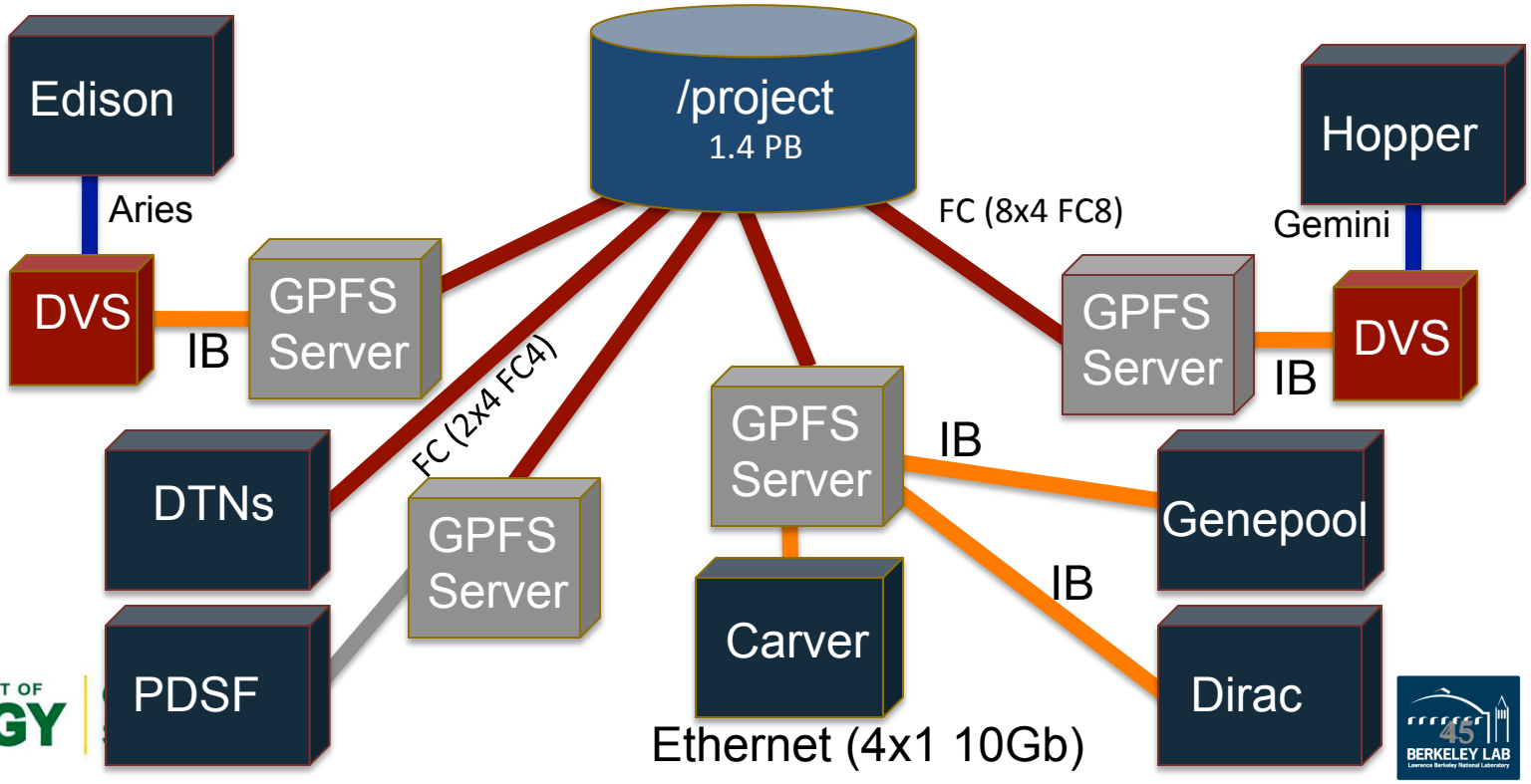
NGF Global Scratch

- **/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



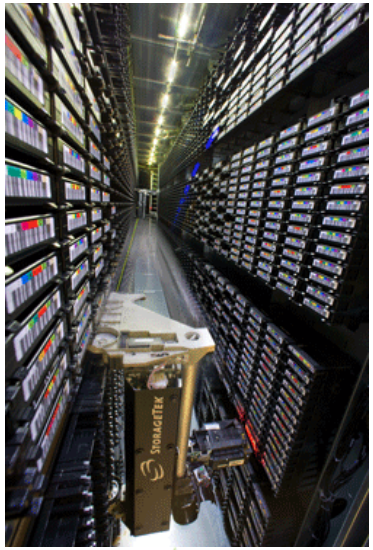
NGF Project

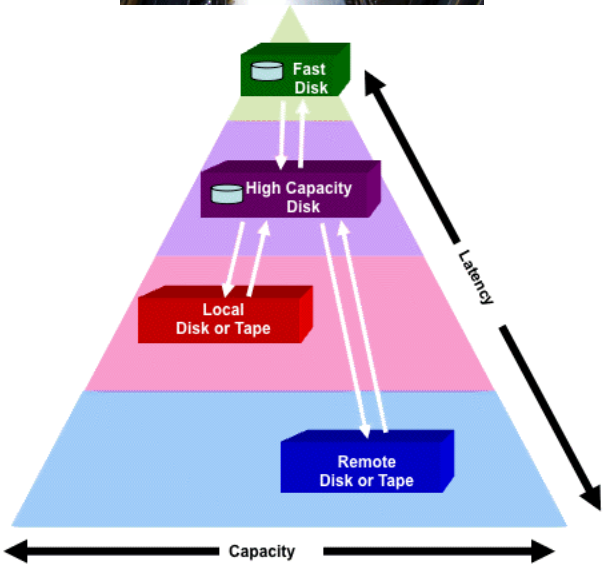
- /project: NERSC-wide sharing and long-term data storage
- Obtain via special request for sharing data between platforms, users, or outside
- Not purged, quota enforced (4TB default per project), backed up daily
- Serves 200 projects; 1.4 PB (+2.8!!) total capacity; ~5 TB average daily IO



Archival Storage: HPSS

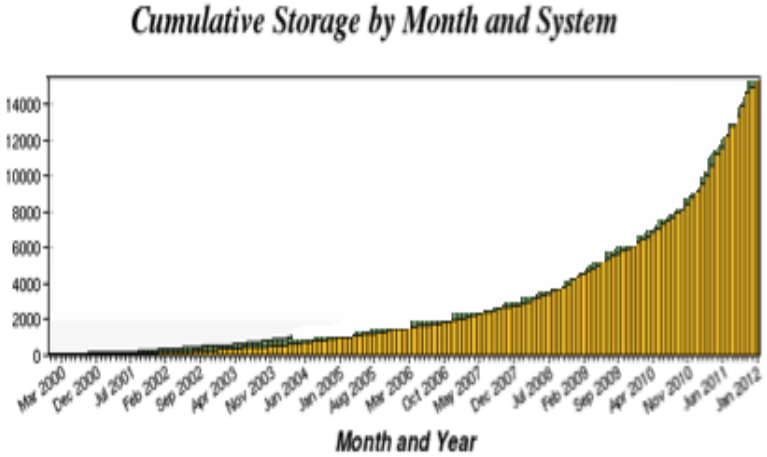
- **For permanent, archival storage**
- **Uses magnetic tape, disk with 150TB fast-access disk cache**
  - ~15 PB data in 140 M files
  - Increases at ~1.7X per year
  - Average data xfer rate: 100 MB/sec
- **Cartridges are loaded unloaded into tape drives by sophisticated robotics**
- **Use HPSS to 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

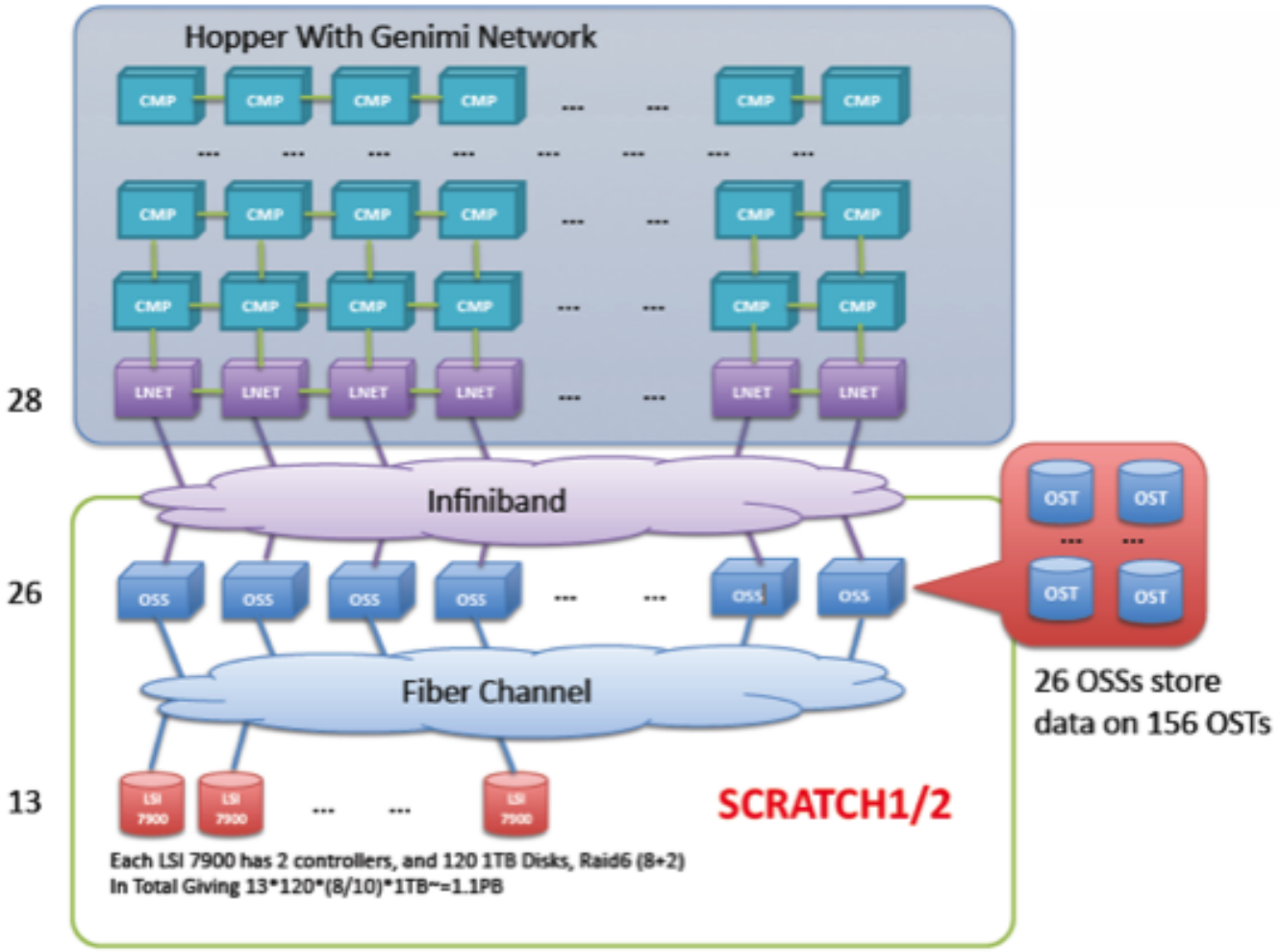
System		Hopper	Edison	Carver	Gene pool	PDSF	Datatrans
Global home	$HOME	✔	✔	✔	✔		✔
Global scratch	$GSCRATCH	✔	✔	✔	✔		✔
Global Project	/project/ projectdirs/ *name*	✔	✔	✔	✔	✔	✔
Local Scratch	$SCRATCH $SCRATCH2	✔	✔•				

- Not all Edison scratch file systems installed yet

Software

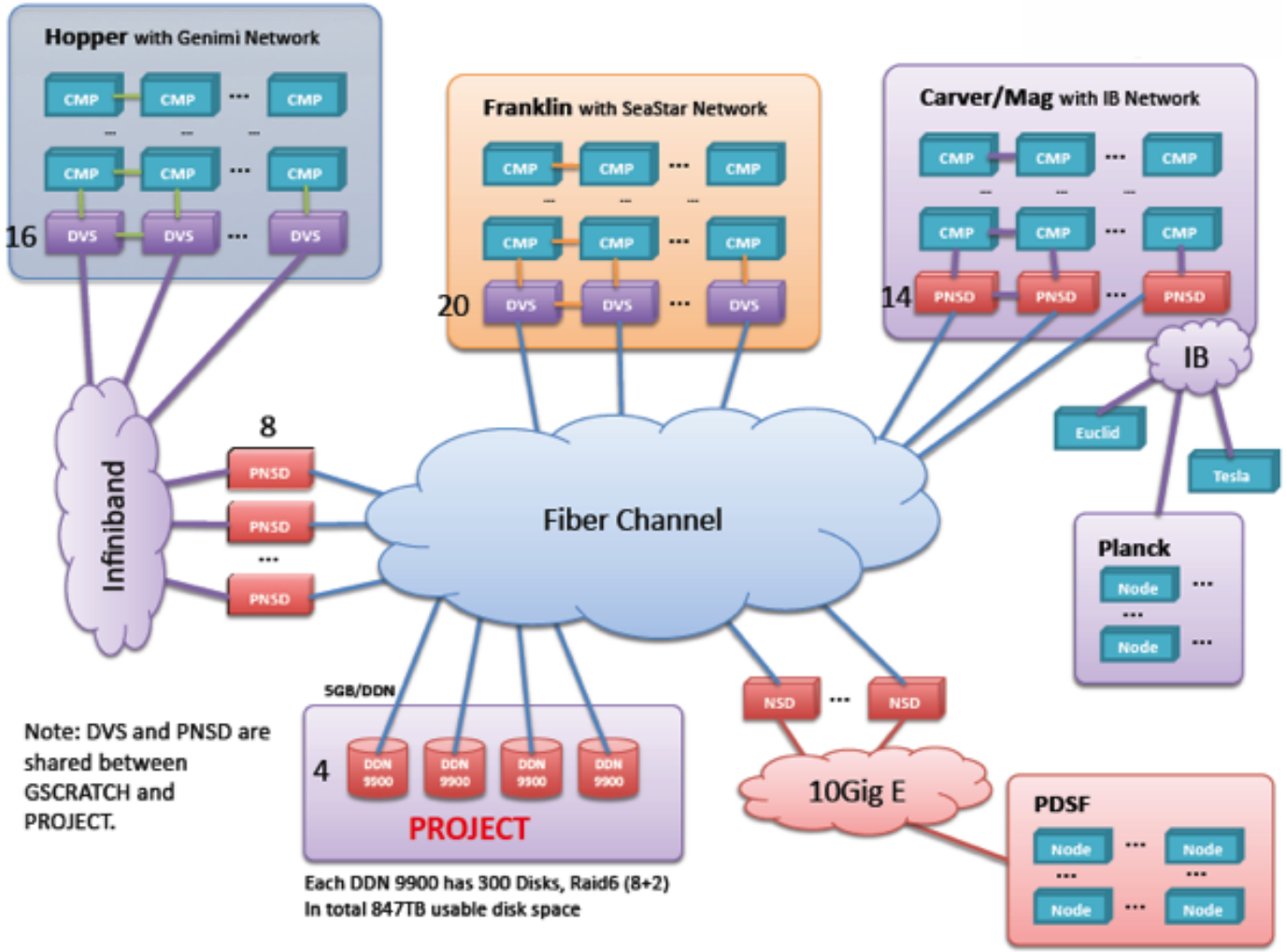
- Vendor supplied
- NERSC supplied
- System supplied
- Requests: consult@NERSC.gov

Hopper Scratch

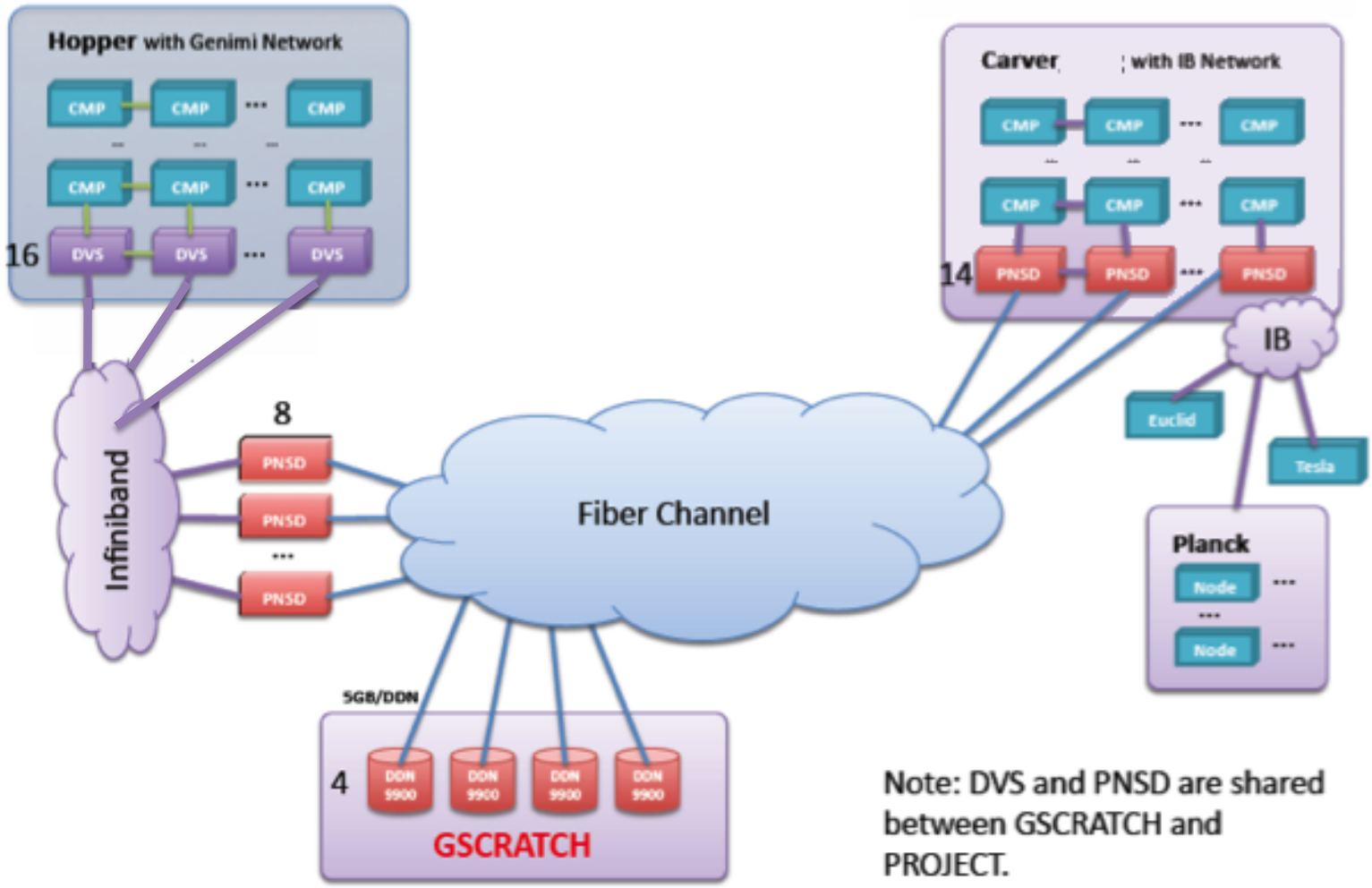


Note: There are two sets of identical configuration for SCRATCH1 and SCRATCH2

/Project


Hopper with Genimi Network
CMP
DVS
16
Franklin with SeaStar Network
20
Carver/Mag with IB Network
PNSD
14
IB
Euclid
Tesla
Planck
Node
Infiniband
8
Fiber Channel
SGB/DDN
4
DDN 9900
PROJECT
NSD
10Gig E
PDSF
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

Global Scratch



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!

Hopper Node Details

- **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.**