Introduction to NERSC

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

January 2012

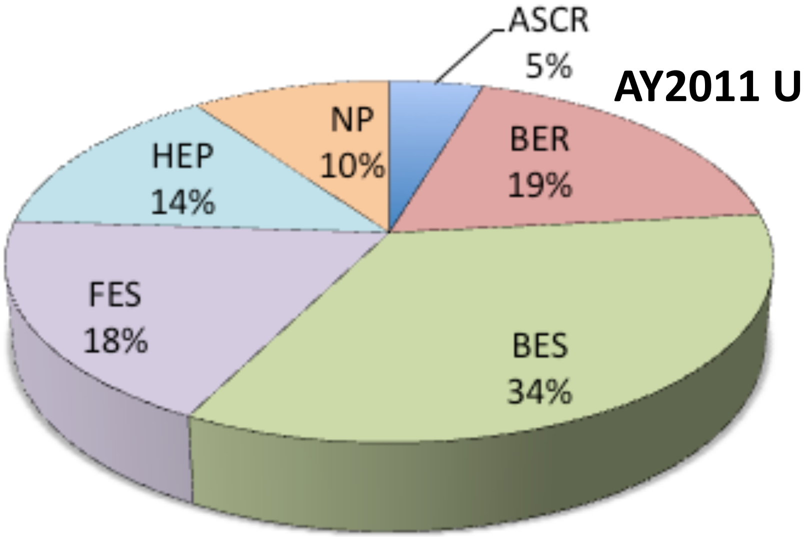
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

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 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



AY2011 Usage

Outline

- Overview of platforms, storage systems
- Usage model
- Miscellaneous

Main NERSC Platforms

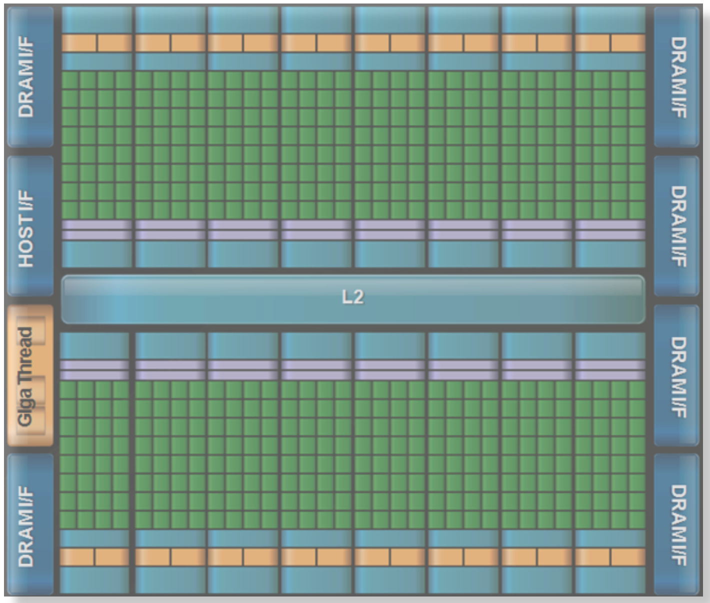
System	Hopper	Franklin	Carver	Euclid
Purpose	Compute	Compute	Compute	Analysis
Nodes	6,384	9,572	1,202	One
Node Contents	2 CPUs X 12 cores	1 X 4	1,120 @ 2 X 4 80 @ 2 X 6	8 X 6
Total Cores	153,216	38,288	9,920	48
CPU	AMD Opteron MagnyCours	AMD Opteron Budapest	Intel Nehalem/ Westmere	AMD Opteron
Memory	**	2 GB/core	**	512 GB Total
Interconnect	Cray "Gemini"	Cray "SeaStar 2+"	4X QDR Infiniband	N/A
Storage ***	2 PB Lustre	0.4 PB Lustre		

Franklin Status

- 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

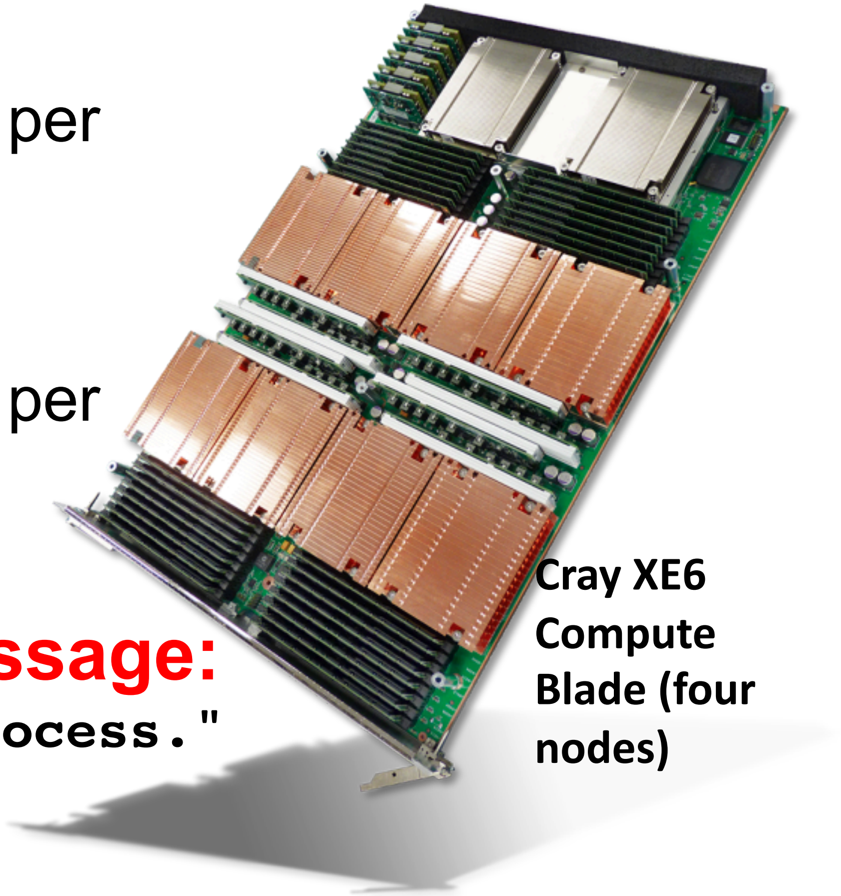
- 50-node “Dirac” GPU test bed
- Data transfer nodes dtn01 and dtn02:
  - Optimize WAN transfer between DOE facilities.
  - Reduce load on computational systems’ login and service nodes
- PDSF



U.S. DEPARTMENT OF ENERGY | Office of Science

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

Hopper & Carver Memory

- David and Richard will tell you how to submit jobs so you can target specific memory configurations.

Hardware Comparisons

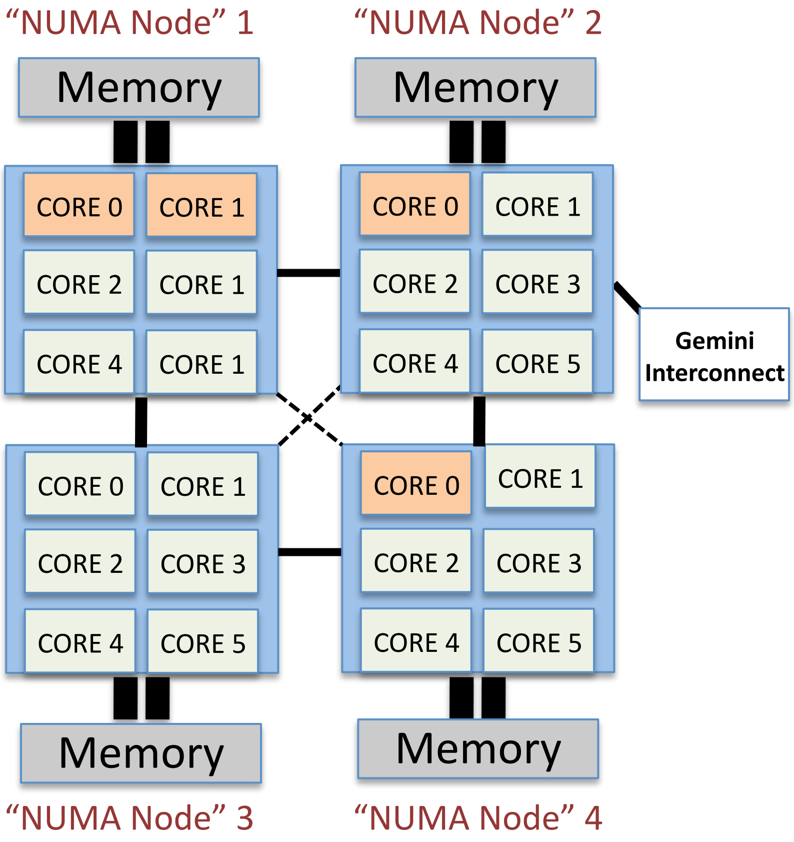
	Clock (GHz)	Cores / Node	Peak GFLOPS / s / node	STREAM GB/s/core			
				PGI	Intel	Cray	GCC
Nehalem	2.6	8	83	4391	4628		
Westmere	2.6	12	125	3298	3516		
Magny-Cours (Hopper)	2.1	24	202	2245	2254	2118	1616
Budapest (Franklin)	2.3	4	37	2298			

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

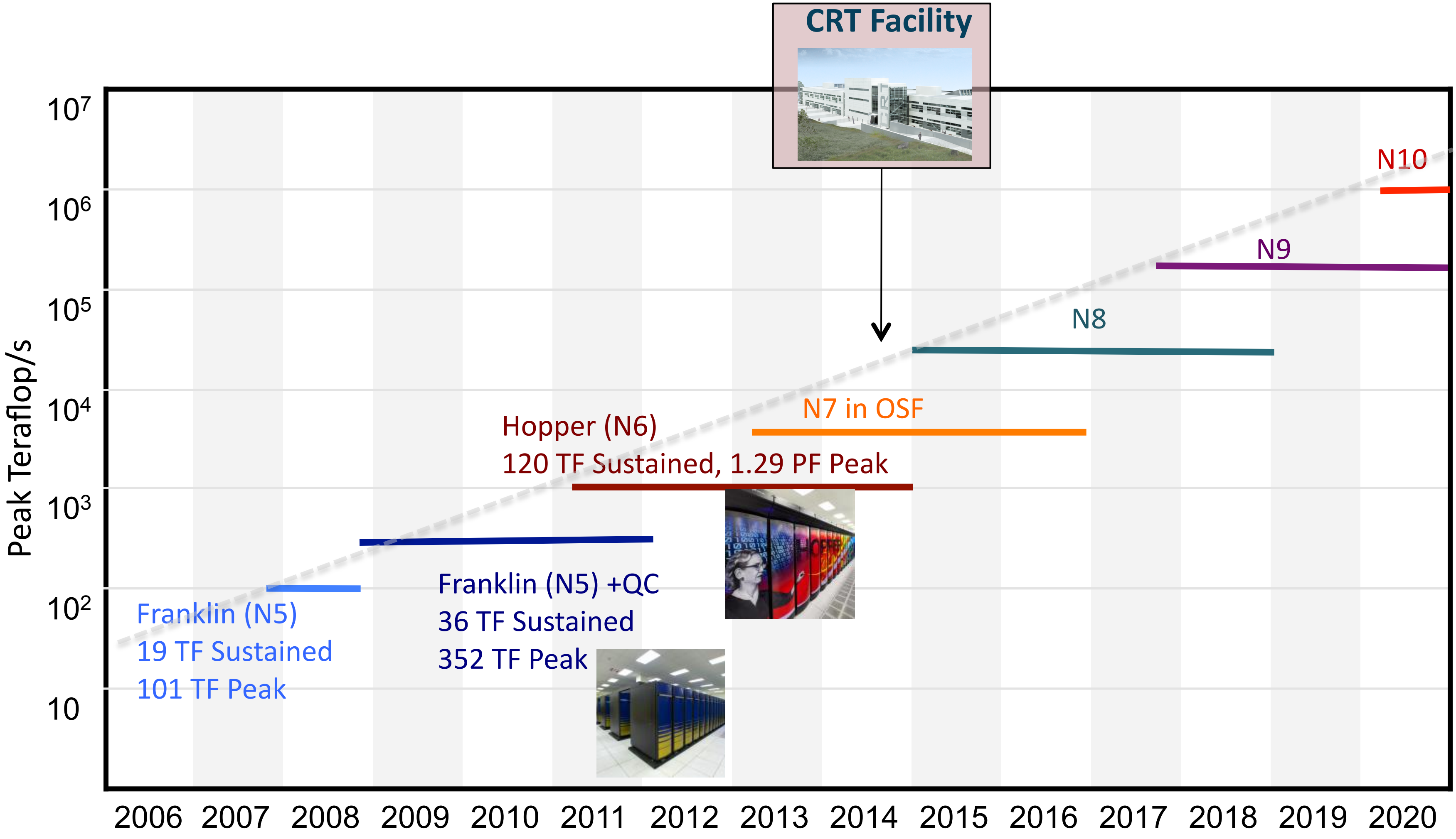
Caution on performance comparisons - 3 different processor generations

Hopper Node Details

- **N**on-**U**niform **M**emory **A**ccess
  - 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.



NERSC Roadmap



We are working on the exact scope for NERSC-7.

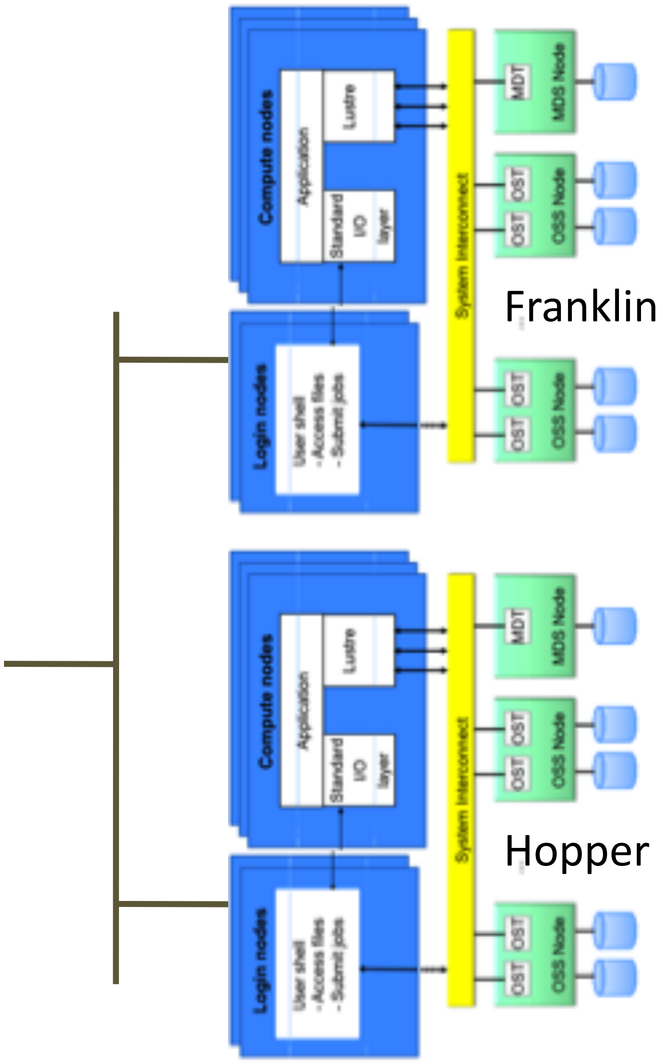
Online Storage Systems

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

- Global file systems

Local 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

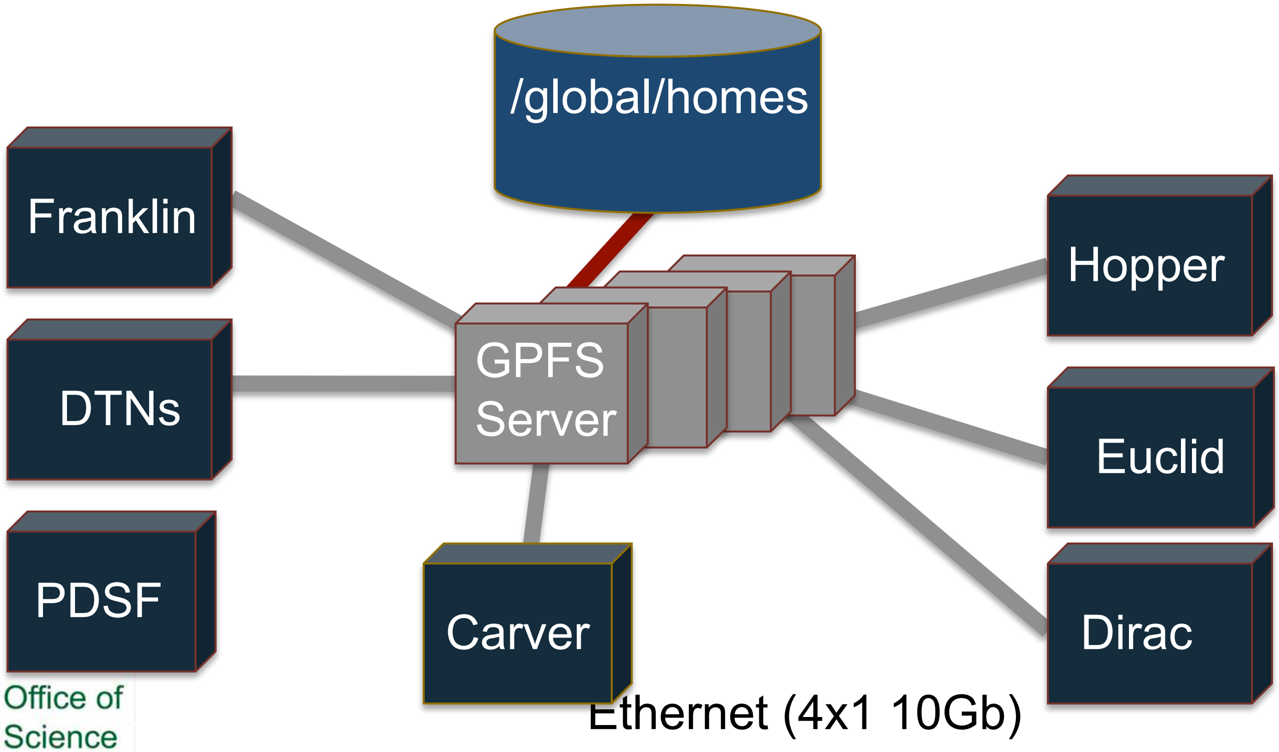
- /global/scratch

- /project

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

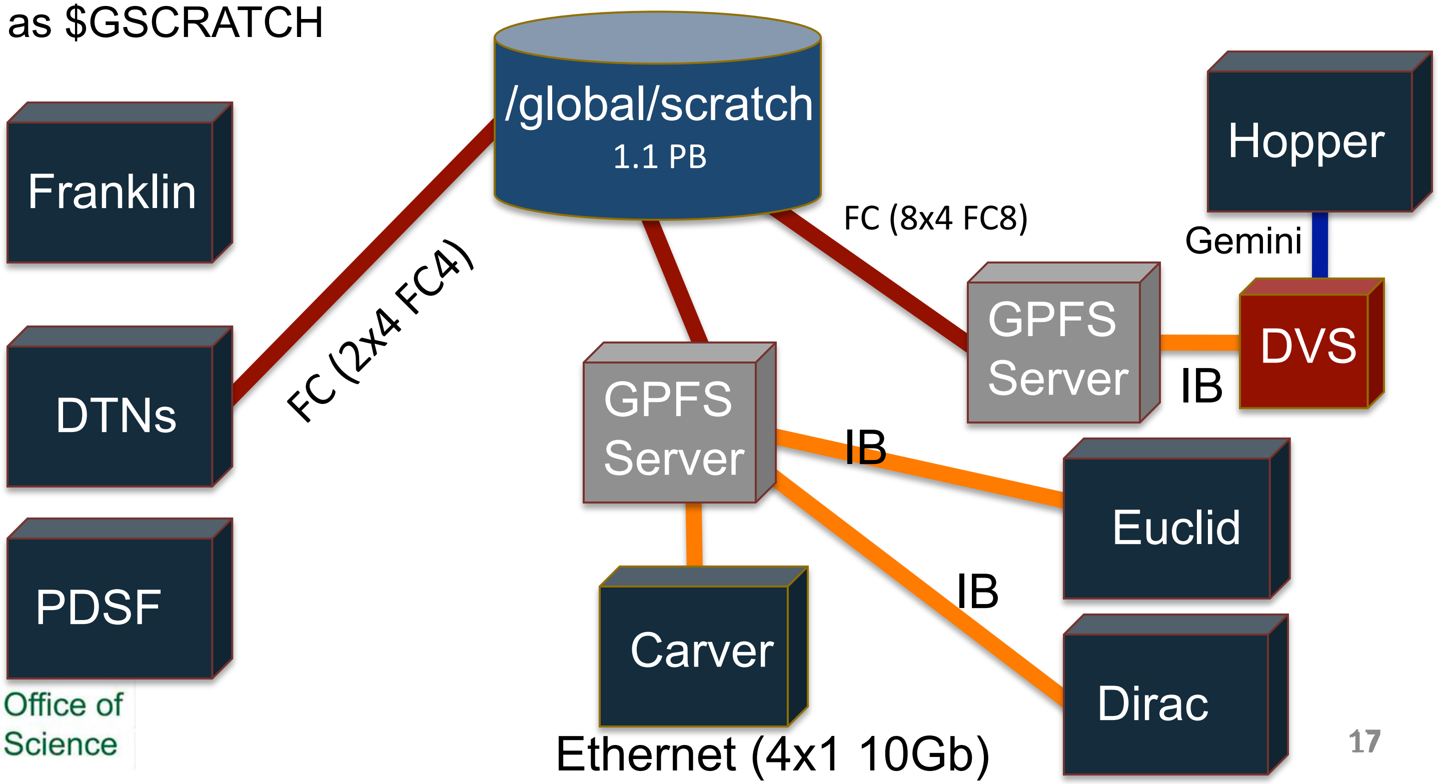
NGF Global Homes

- **/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.



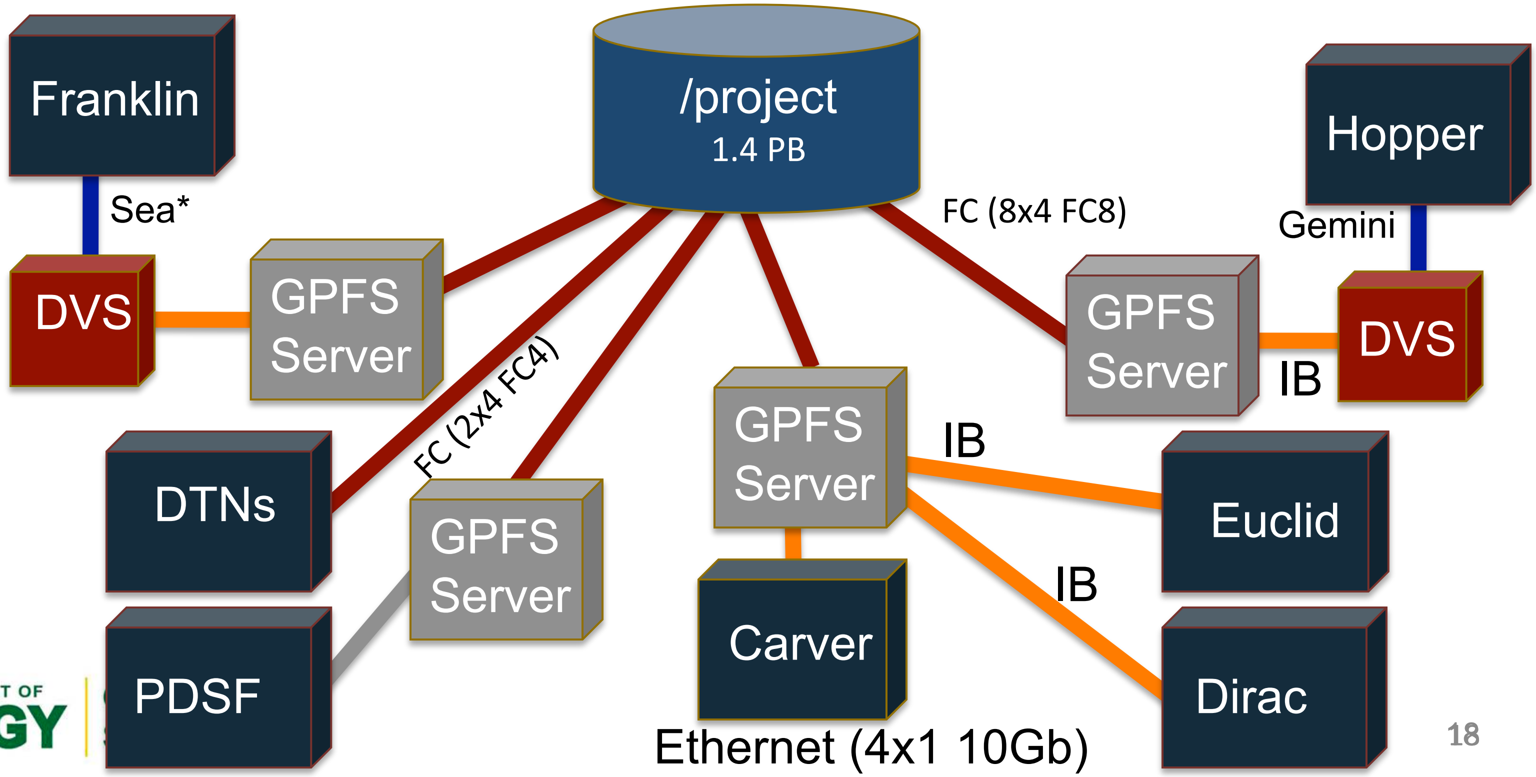
NGF Global Scratch

- **/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



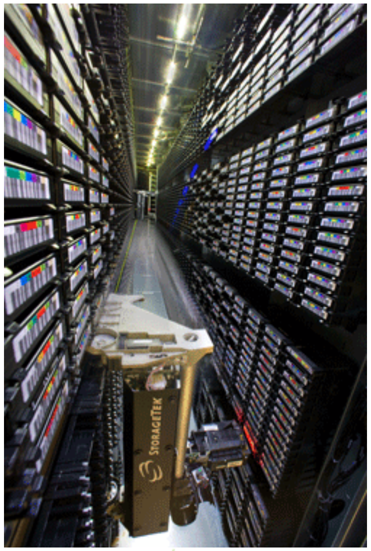
NGF Project

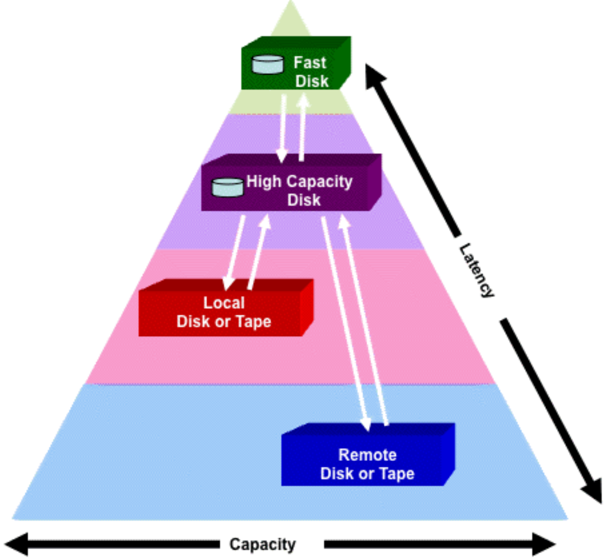
- **/project: NERSC-wide sharing and long-term data storage**
- Obtain via special request for sharing data between platforms, users, or outside
- Not purged, quota enforced (4TB default per project), backed up daily
- Serves 200 projects; 1.4 PB (+2.8!!) total capacity; ~5 TB average daily IO



Archival Storage: HPSS

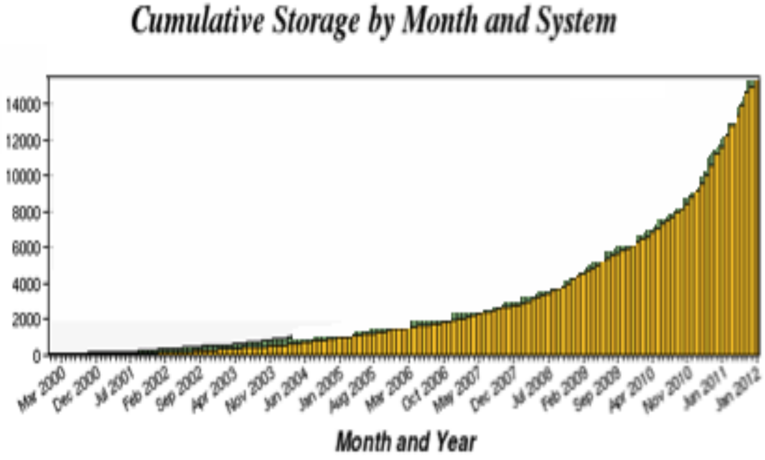
- For permanent, archival storage
- Uses magnetic tape, disk with 150TB fast-access disk cache
  - ~15 PB data in 140 M files
  - Increases at ~1.7X per year
  - Average data xfer rate: 100 MB/sec
- Cartridges are loaded unloaded into tape drives by sophisticated robotics
- Use HPSS to back up your code,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



Cumulative Storage by Month and System

* clients available for download

** not ssh

Usage Model

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

Login Nodes

- 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

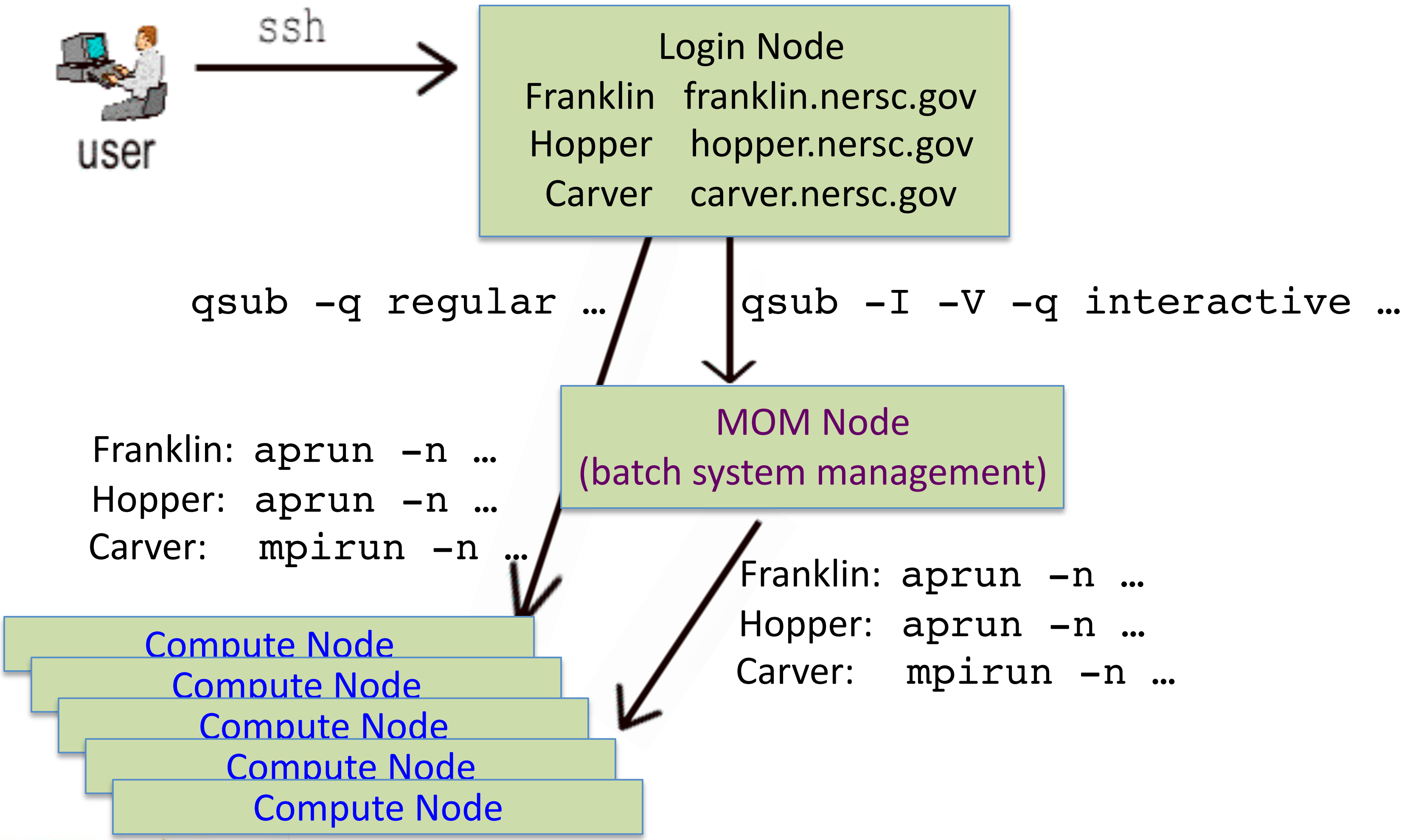
Compute Nodes

- 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

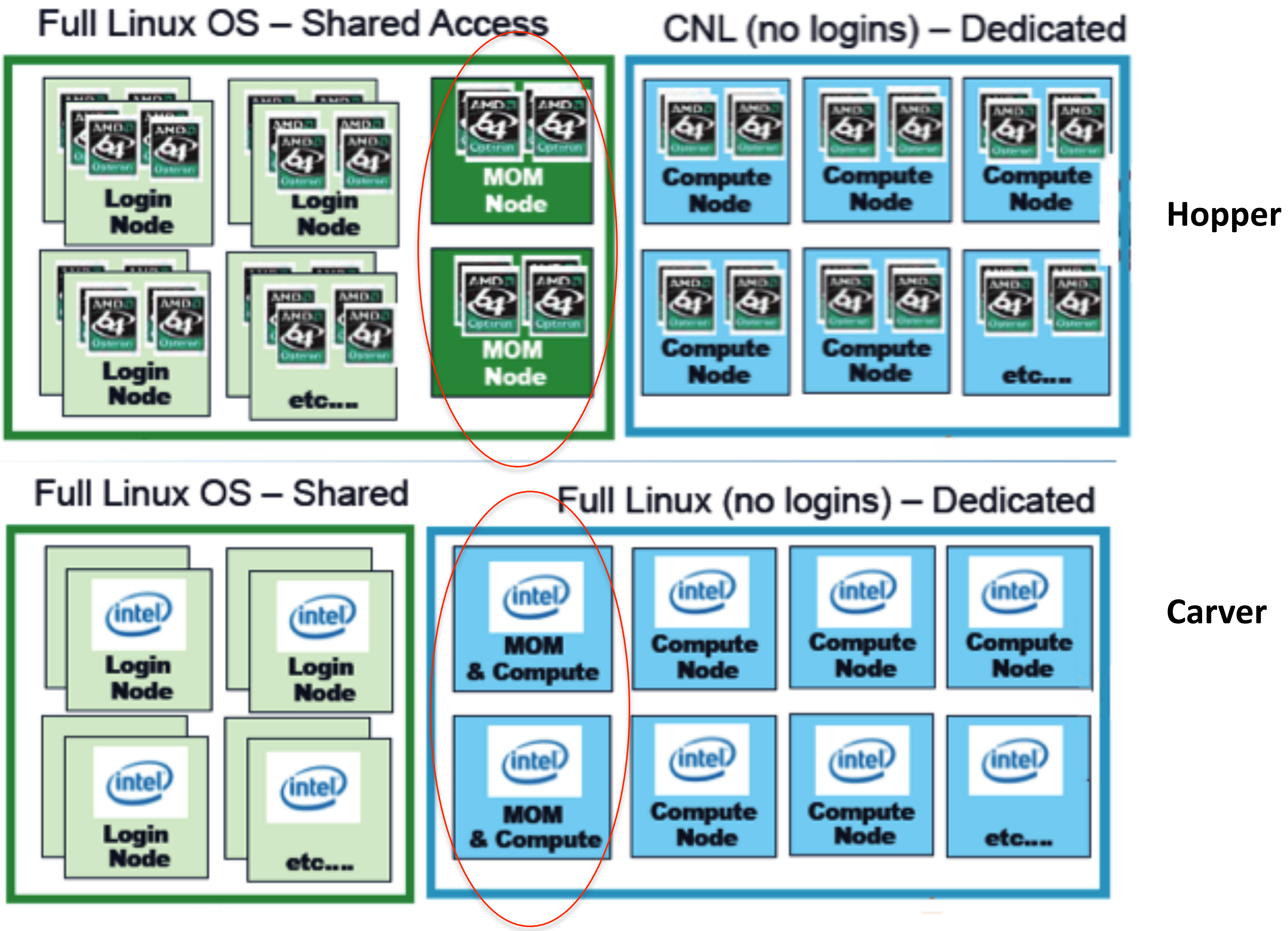
Service Nodes

- “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



Service Node Configuration

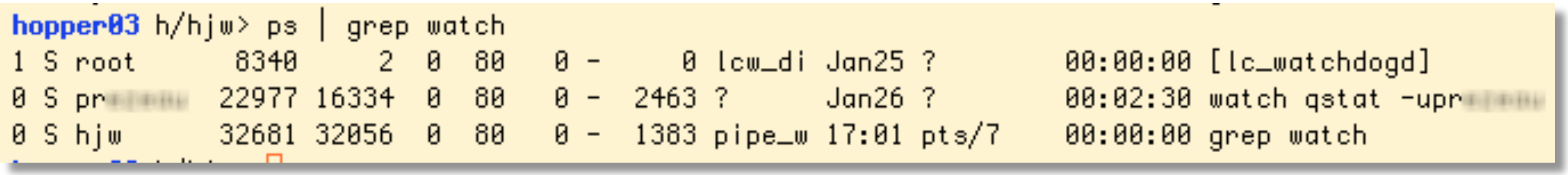


Choosing a System

- 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

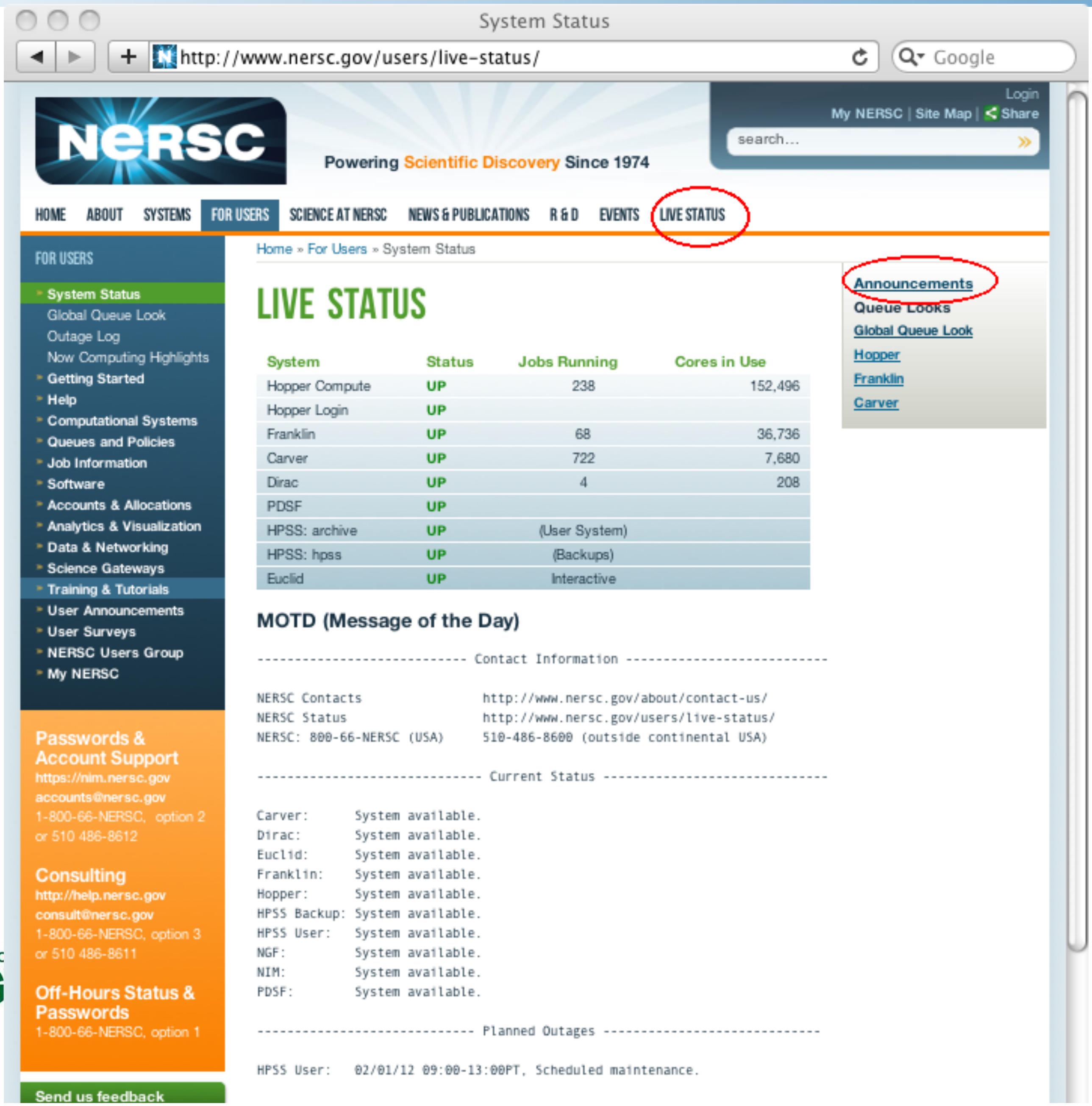
Important Policies

- 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 pr[redacted] 22977 16334  0  80   0 -  2463 ?      Jan26 ?        00:02:30 watch qstat -upr[redacted]
0 S hjw       32681 32056  0  80   0 -  1383 pipe_w 17:01 pts/7    00:00:00 grep watch
```

Important Web Page



LIVE STATUS

System	Status	Jobs Running	Cores in Use
Hopper Compute	UP	238	152,496
Hopper Login	UP		
Franklin	UP	68	36,736
Carver	UP	722	7,680
Dirac	UP	4	208
PDSF	UP		
HPSS: archive	UP	(User System)	
HPSS: hpss	UP	(Backups)	
Euclid	UP	Interactive	

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: 800-66-NERSC (USA)   510-486-8600 (outside continental USA)

----------------------------- Current Status ------------------------------

Carver:      System available.
Dirac:       System available.
Euclid:      System available.
Franklin:    System available.
Hopper:      System available.
HPSS Backup: System available.
HPSS User:   System available.
NGF:         System available.
NIM:         System available.
PDSF:        System available.

---------------------------- Planned Outages -----------------------------

HPSS User:   02/01/12 09:00-13:00PT, Scheduled maintenance.
```

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

Off-Hours Status & Passwords
1-800-66-NERSC, option 1

Getting Help

http://www.nersc.gov

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

Computer Operations* = menu option 1 (24/7)

Account Support = menu option 2

accounts@nersc.gov

HPC Consulting = menu option 3

consult@nersc.gov

(8-5, M-F Pacific time)

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

* Passwords during non-business hours

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 (and talks).
- 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/

The Journal of Chemical Physics
nature
JACS
Applied Physics Letters
Structure
Journal of Fluid Mechanics
C&EN
Journal of Computational Chemistry
Physics of Plasmas

1500 publications per year

Thank you.

Additional Info

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

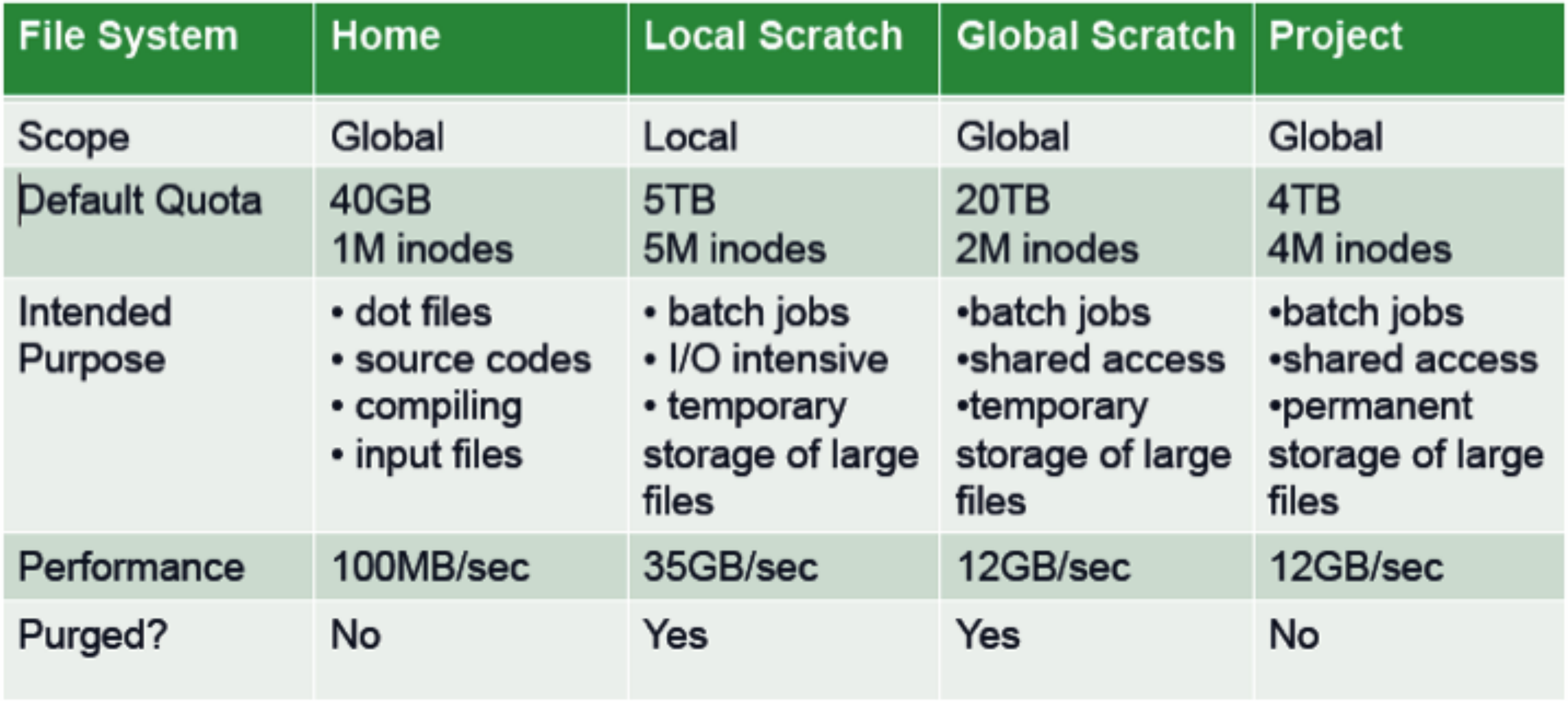
"Leadership Facilities" at Oak Ridge & Argonne

- 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

File System Availability

System		Hopper	Franklin	Carver	Euclid	PDSF	Datatrans
Global home	$HOME	✓	✓	✓	✓		✓
Global scratch	$GSCRATCH	✓		✓	✓		✓
Global Project	/project/ projectdirs/ *name*	✓	✓	✓	✓	✓	✓
Local Scratch	$SCRATCH $SCRATCH2	✓	✓				

File System Summary



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

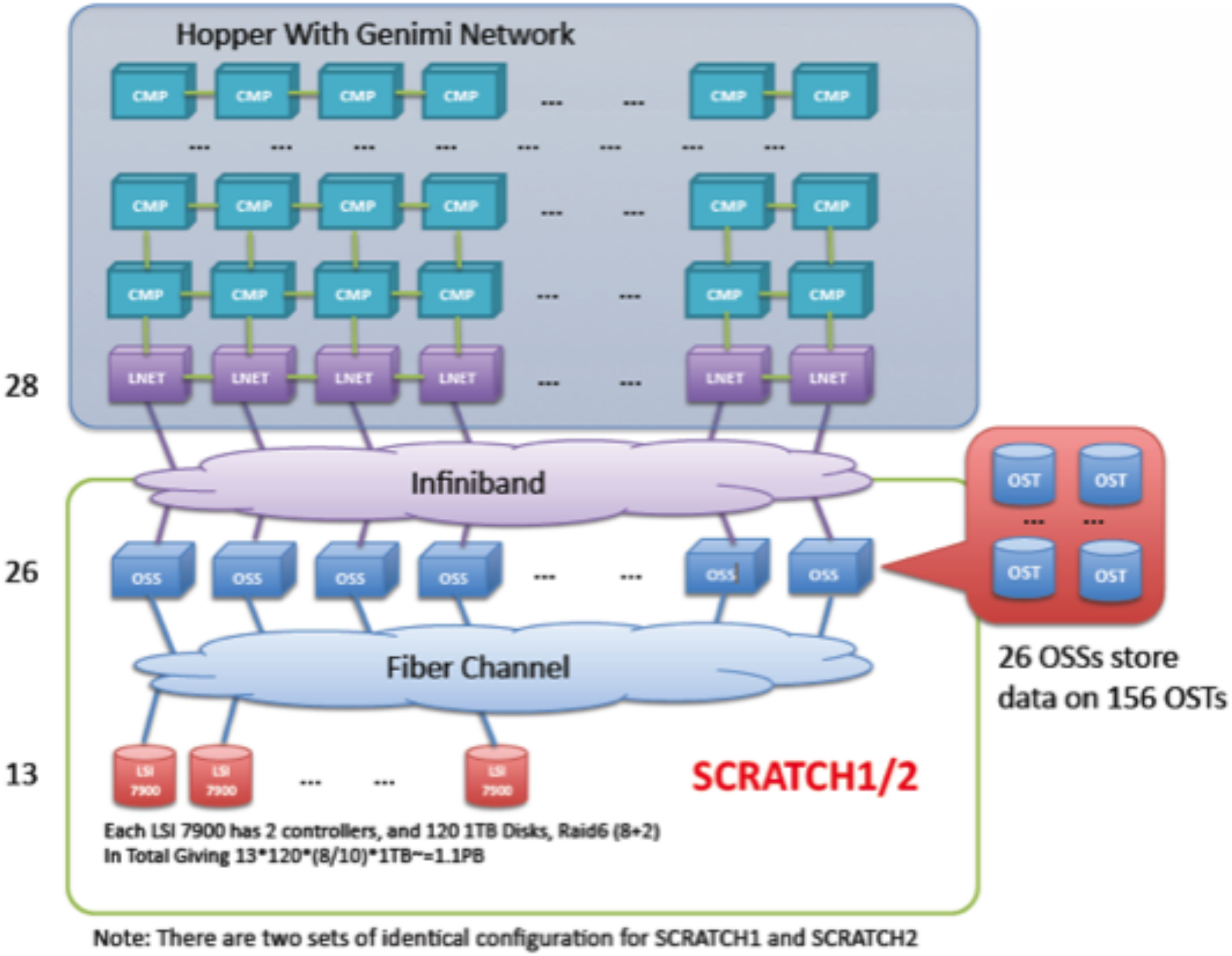
Software

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

DVS

- 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

Hopper Scratch

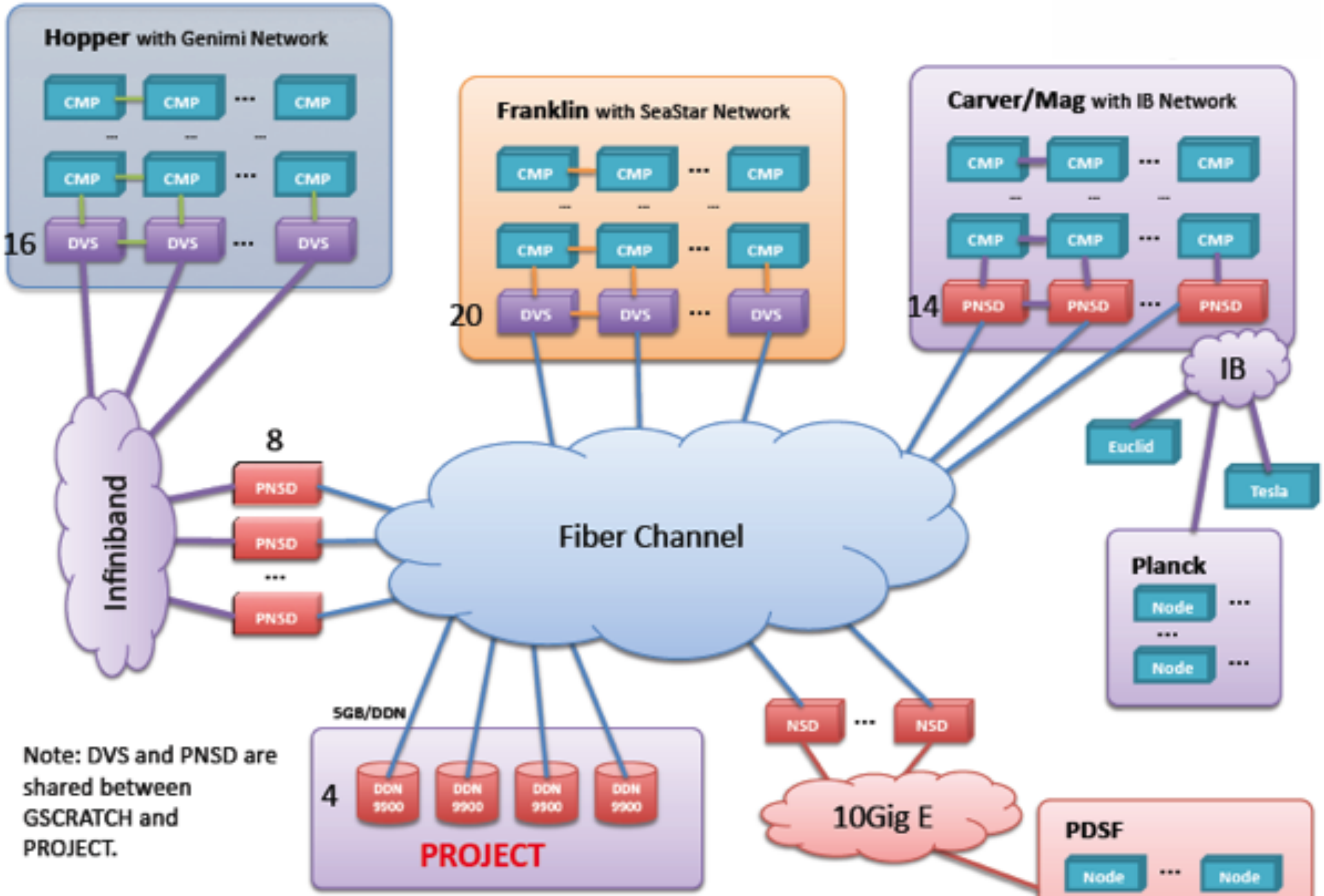


26 OSSs store data on 156 OSTs

Each LSI 7900 has 2 controllers, and 120 1TB Disks, Raid6 (8+2)
In Total Giving 13*120*(8/10)*1TB~=1.1PB

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

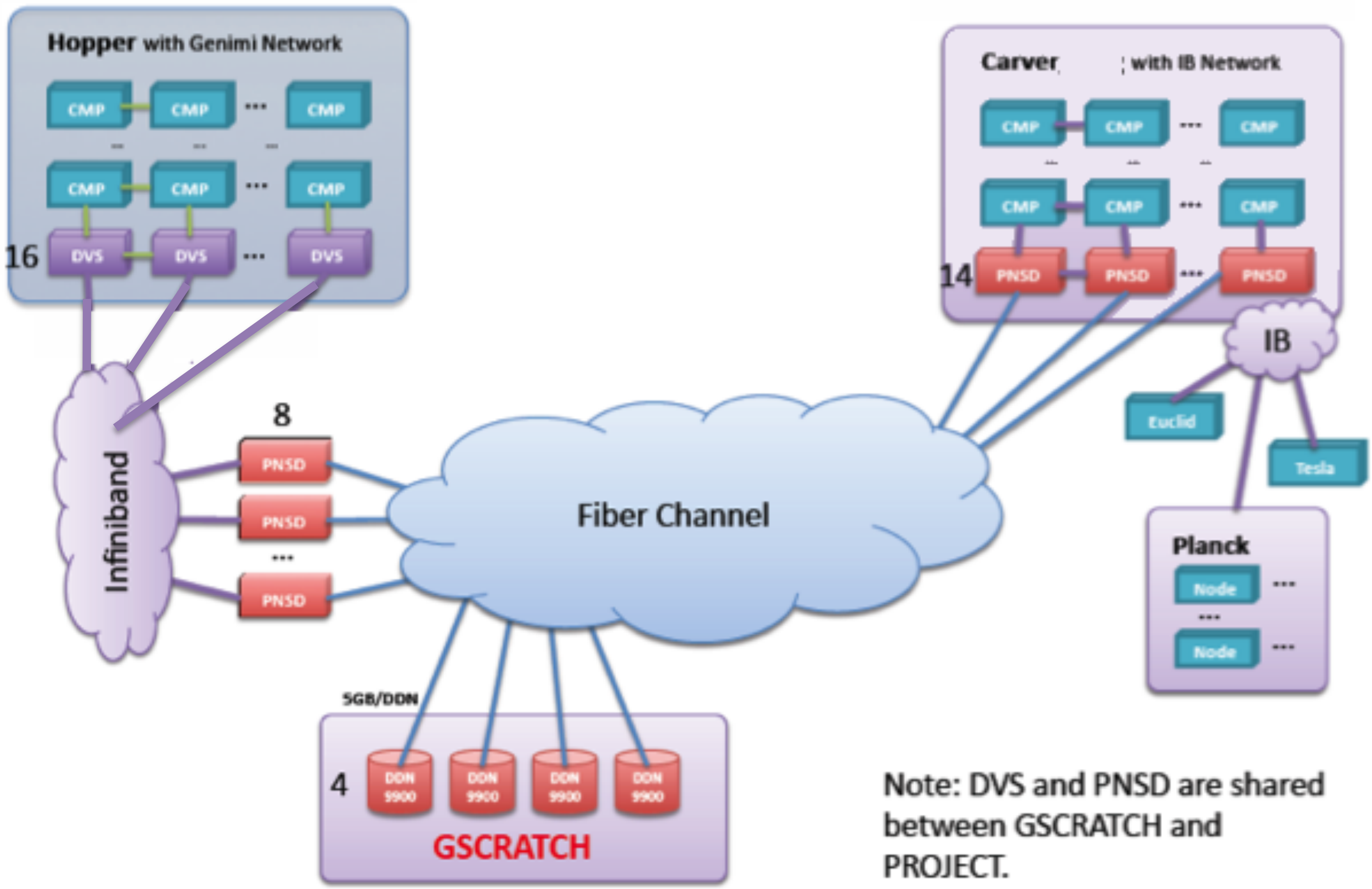
/Project



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

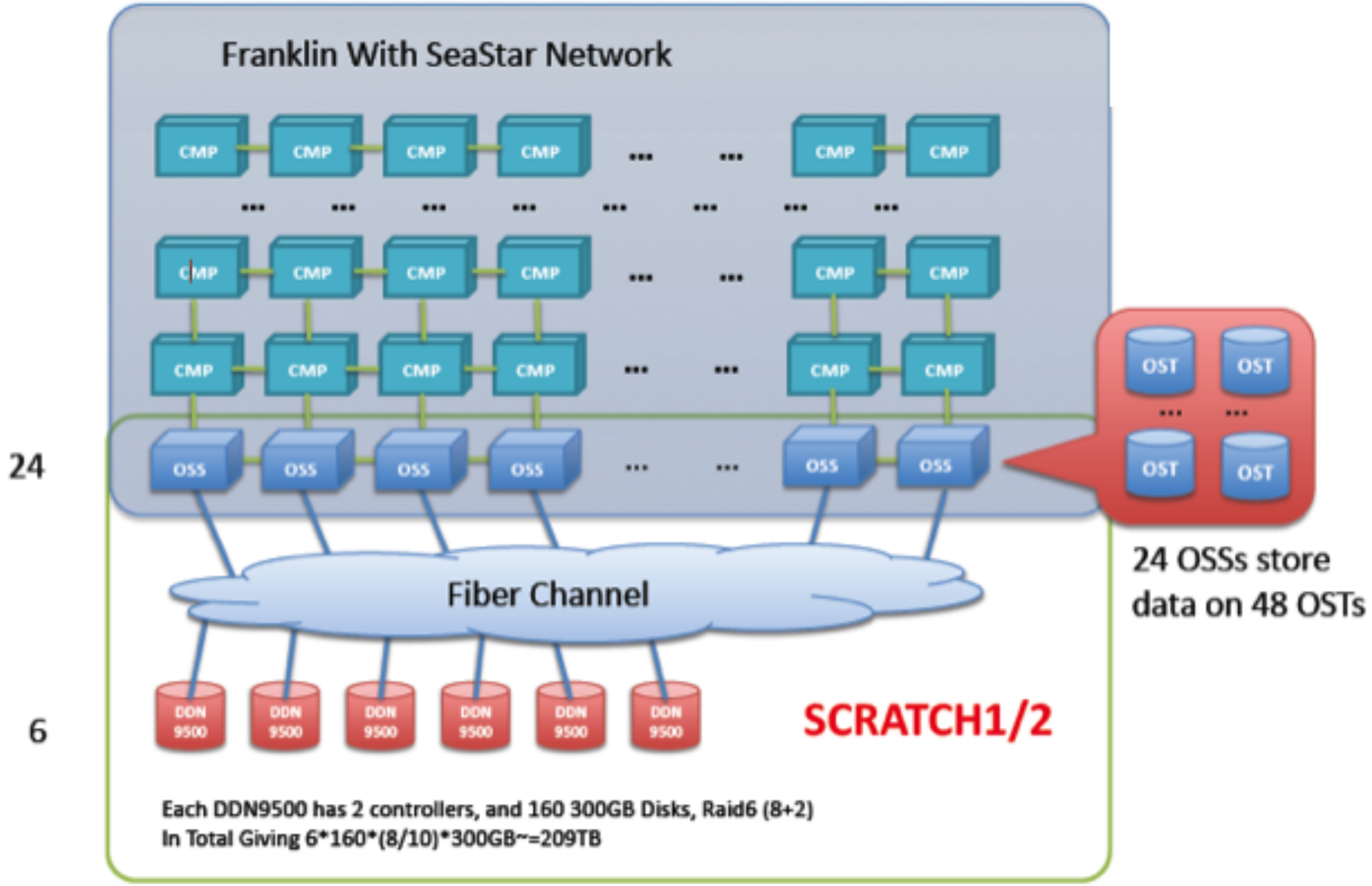
Global Scratch



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

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

Franklin Scratch



24 OSSs store data on 48 OSTs

SCRATCH1/2

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

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!