Introduction to NERSC

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

Harvey Wasserman, NERSC User Services Group

http://www.nersc.gov/users/training/

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

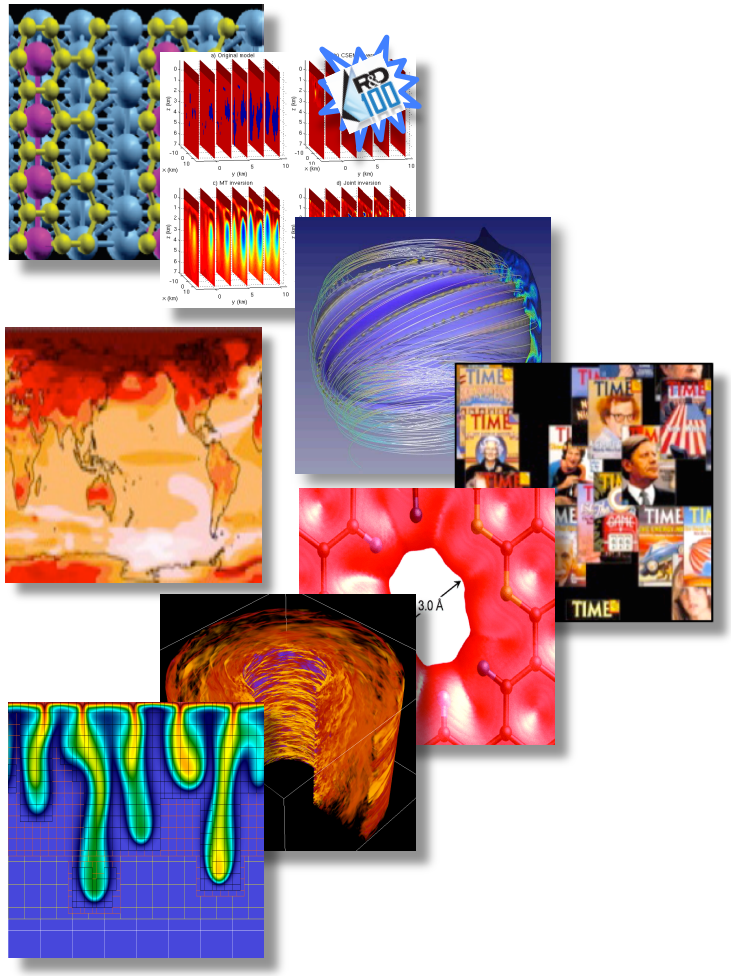
- Established 1974, first unclassified supercomputer center
- Original mission: computational science to complement a magnetically controlled plasma experiment
- Today's mission: ***accelerate the pace of scientific discovery*** by providing production high-performance computing, data, and communications services for research sponsored by the six DOE Office of Science offices

1500 publications per year

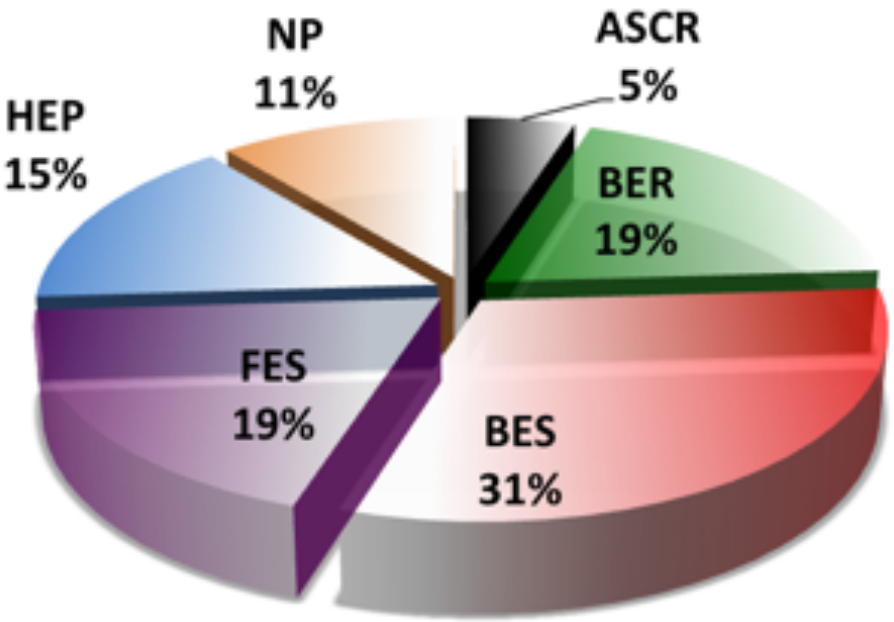
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC is the Production Facility for DOE Office of Science

- Diverse workload:
  - 4,000 users, 550 projects,
  - 700 codes
- 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")



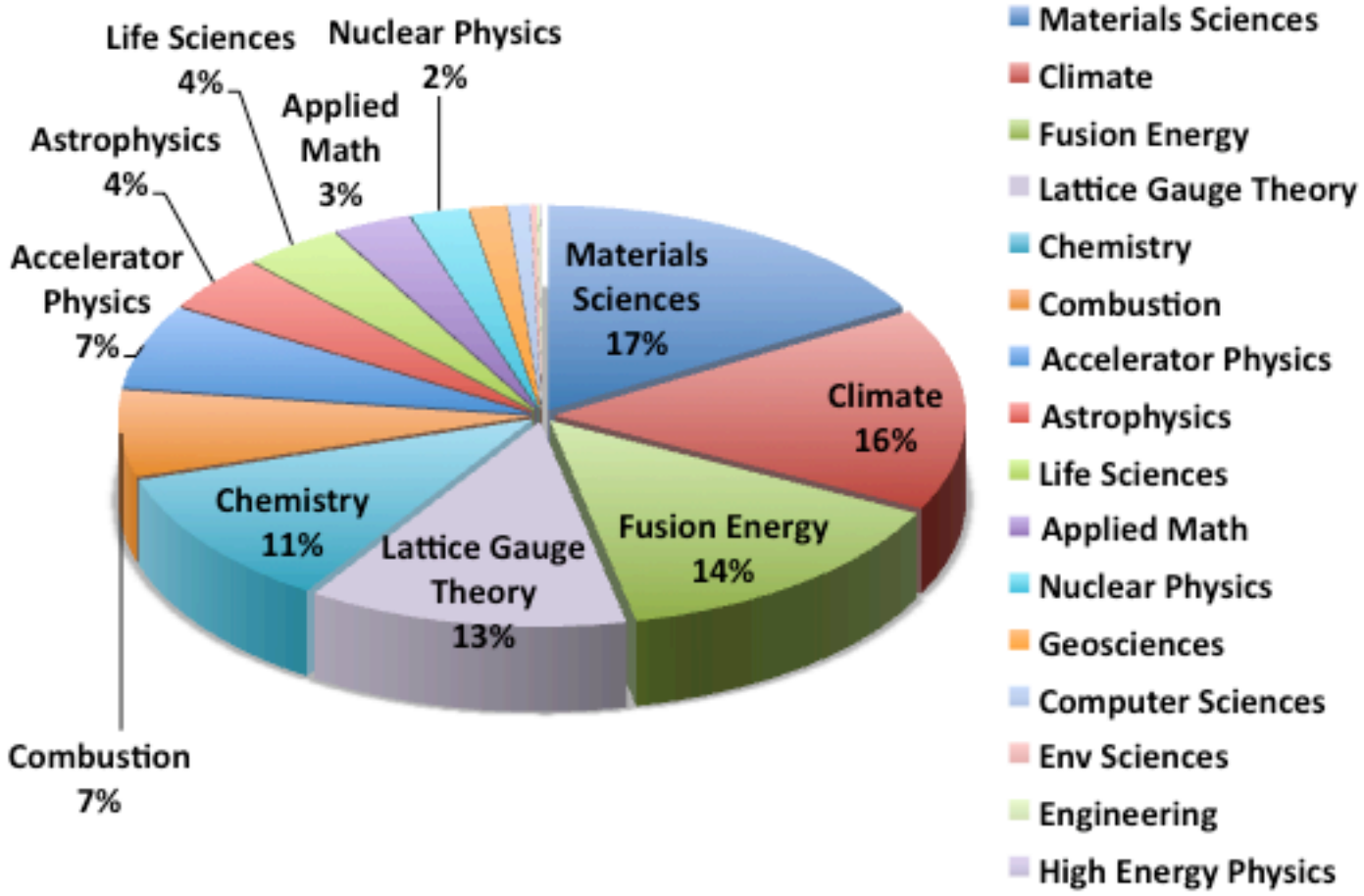
DOE View of Workload



NERSC 2010 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 2008 Allocations
By Science Area

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

U.S. DEPARTMENT OF ENERGY | Office of Science

Main NERSC Platforms

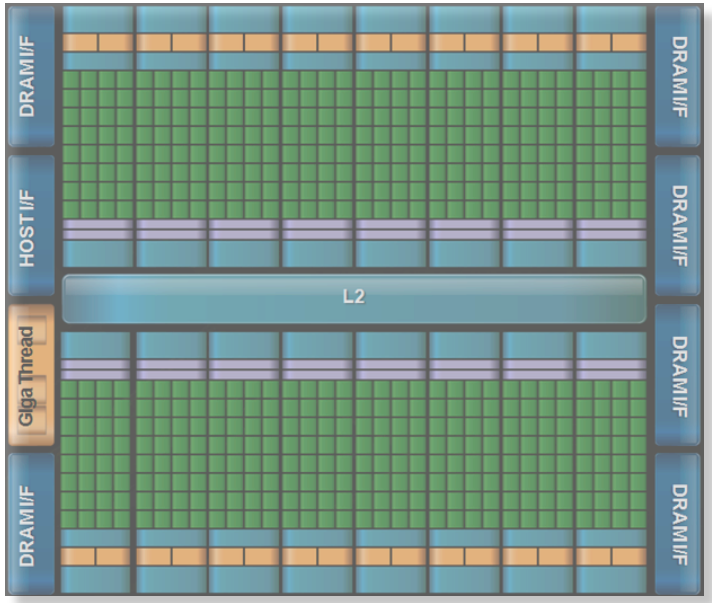
System	Hopper	Carver
Purpose	Compute	Compute
Nodes	6,384	1,202
Node Contents	2 CPUs X 12 cores	1,120 @ 2 X 4 80 @ 2 X 6
Total Cores	153,216	9,920
CPU	AMD Opteron MagnyCours	Intel Nehalem/ Westmere
Memory	32GB/node	24GB/node
Interconnect	Cray "Gemini"	4X QDR Infiniband
Storage ***	2 PB Lustre	

Hopper

Carver

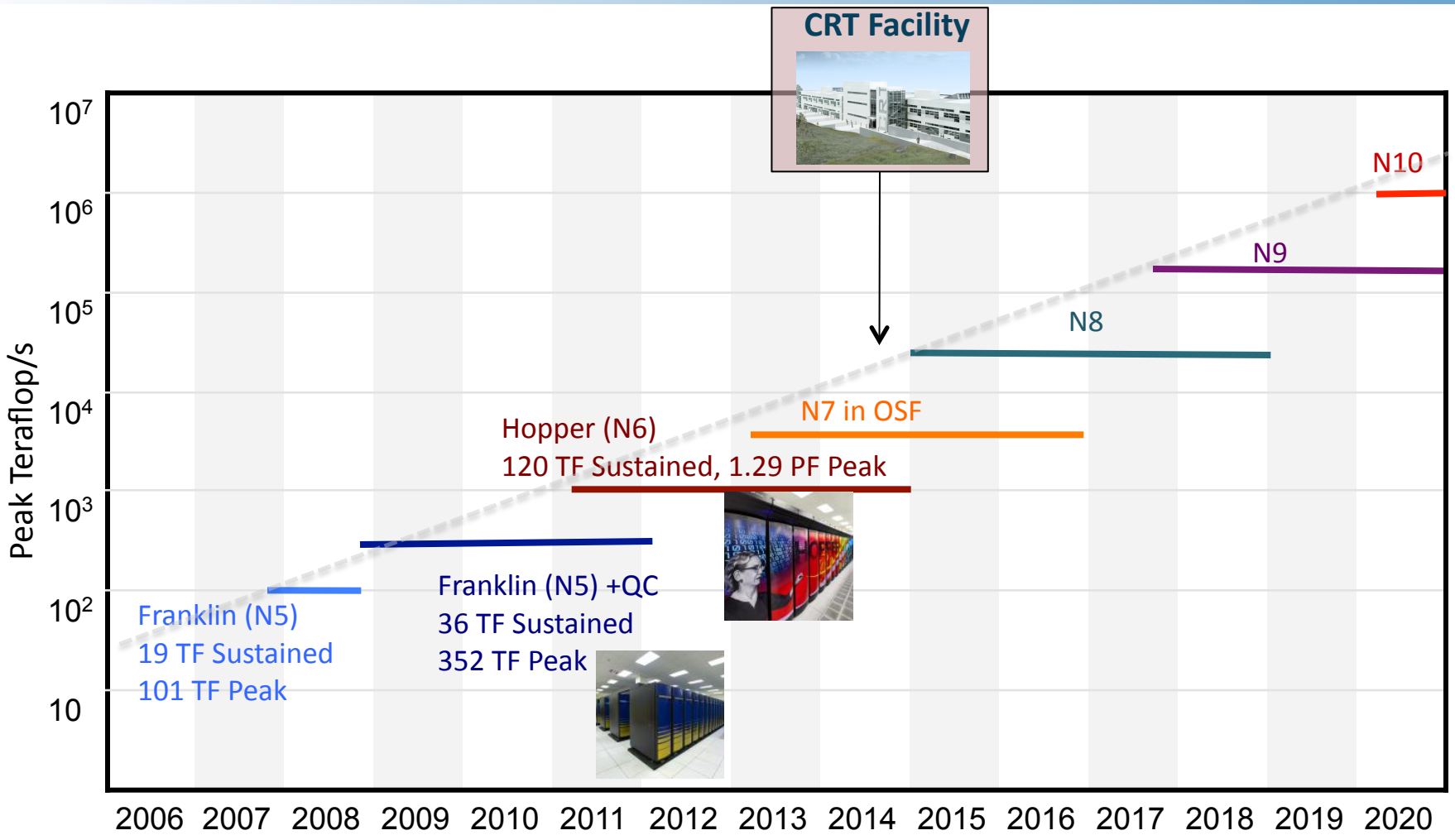
Other NERSC Systems

- 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



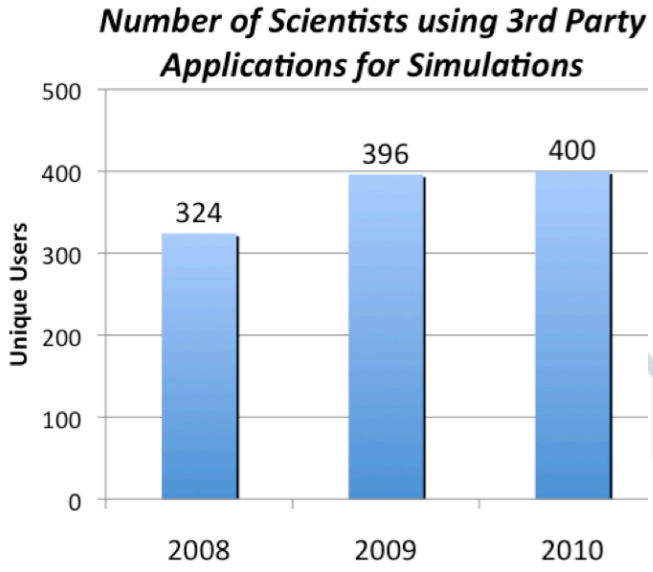
U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC Roadmap



We are working on the exact scope for NERSC-8.

Chemistry & Materials Applications



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

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**

Outline

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

Logging In

```
term% ssh –Y train#@carver.nersc.gov
term% ssh –l train1 –Y carver.nersc.gov
```

Customizations go in:

.cshrc.ext

or

.bashrc.ext

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`** ***`modulefile_name1 modulefile_name2`***
- See what's included:
  - **`module show modulefile_name`**
- Some basic help:
  - **`module help`** ***`modulefile_name`***

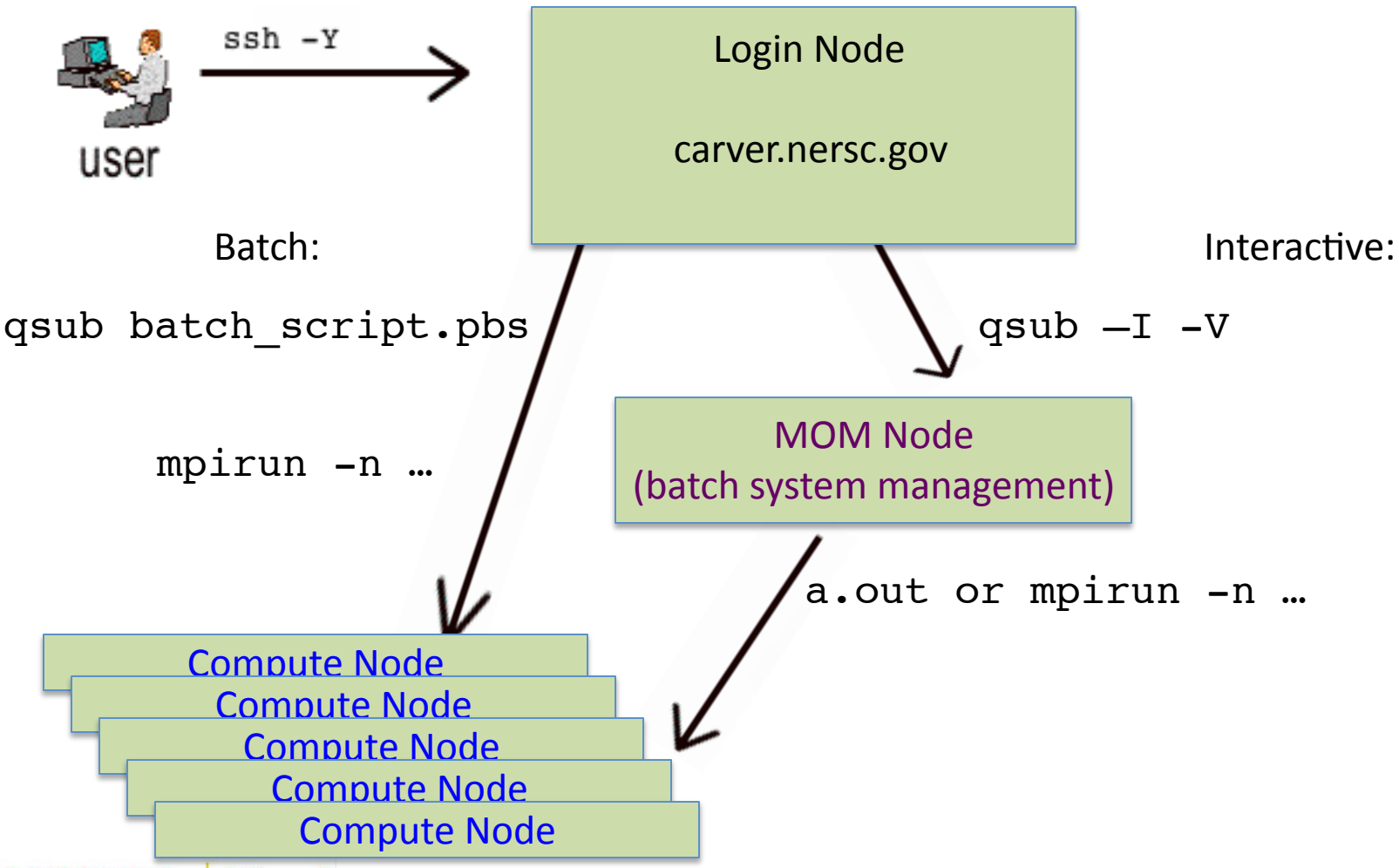
Compiling Environments

- Compiling environments: PGI, Intel, GNU
  - Default is PGI on all NERSC systems
  - 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
  - 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; 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`**

Running Jobs



Notes

- **`qsub –I –V`** has 30-minute time limit
  - Use **`qstat –u *userID*`** to show status
- **`qsub –I –V`** puts you in your $HOME directory
  - Get back to submit dir: **`cd $PBS_O_WORKDIR`**
- **`qsub –I –V gives you 8 processors`**
  - To get more: **`qsub –I –V –l nodes=2:ppn=8`**
- **`qsub –I –V`** uses the “debug” queue
  - One job per user allowed at a time – please remember to “exit”
- Use login node for compiling, editing; use compute node for running programs – access via batch system only

Sample Batch Script

```
#PBS -q regular
#PBS -l nodes=1:ppn=8
#PBS -l walltime=00:35:00
#PBS -N my_job
#PBS -j oe
#PBS -V

cd $PBS_O_WORKDIR
mpirun –np 8 ./petsc_ex2
```

Thank you.

Additional Info

Usage Model

- Compute nodes run applications.
- Service nodes handle support functions: reach via the batch system
- Login nodes provide additional user services: reach via **`ssh –Y carver.nersc.gov`**

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

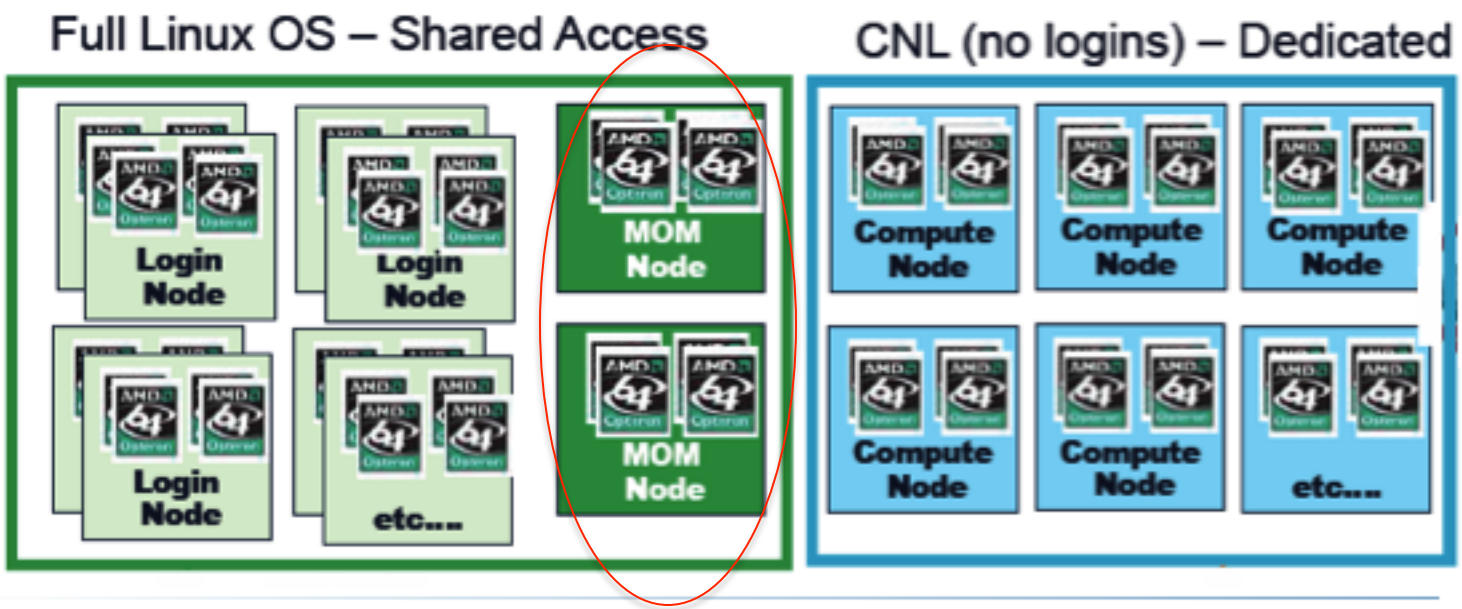
Compute Nodes

- 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: Hopper < Carver

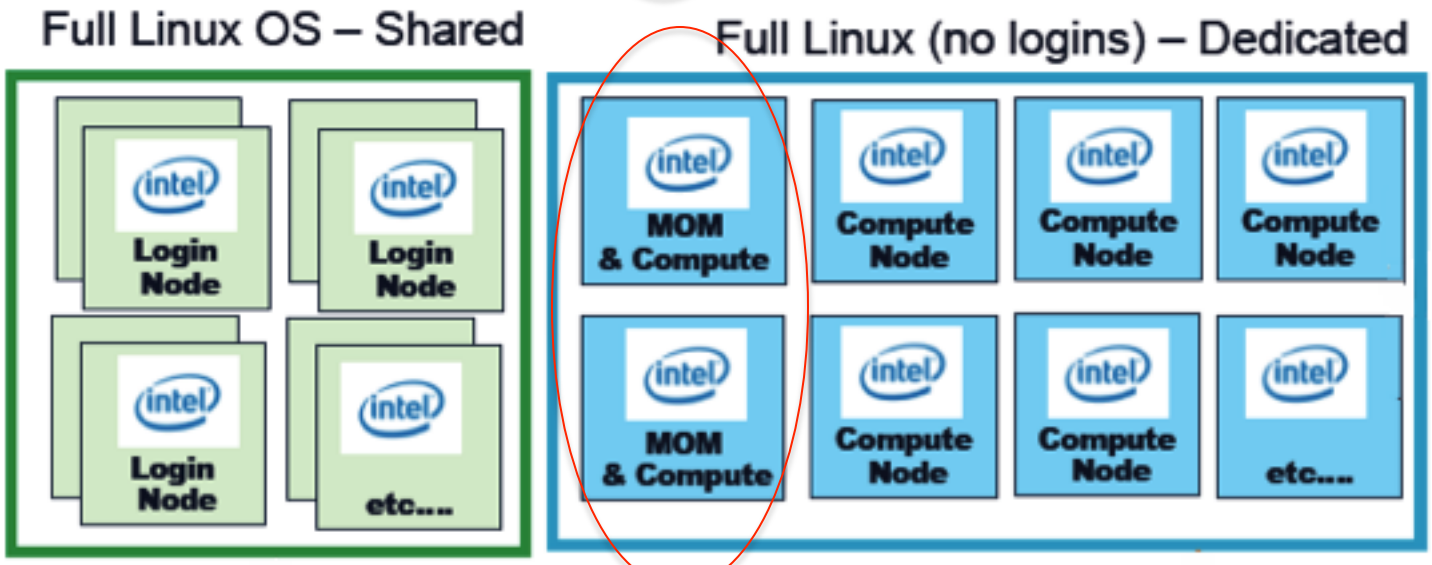
Service Nodes

- "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 Hopper
- Hopper: separate node; Carver: compute node
- Keeping the load down is imperative

Service Node Configuration

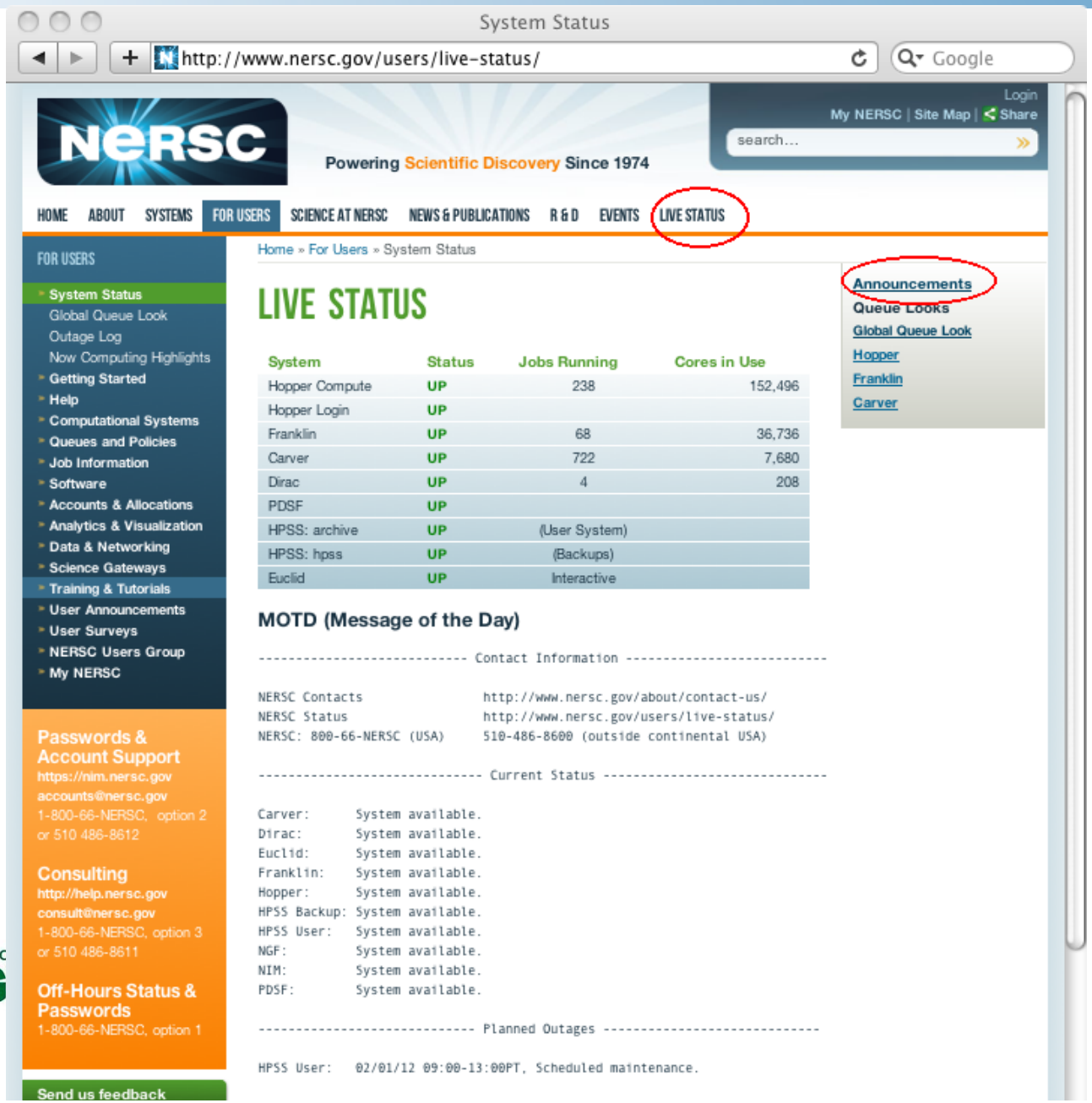


Hopper



Carver

Important Web Page



U.S. DEPARTMENT OF ENERGY

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

U.S. DEPARTMENT OF ENERGY | Office of Science

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.

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

Online Storage Systems

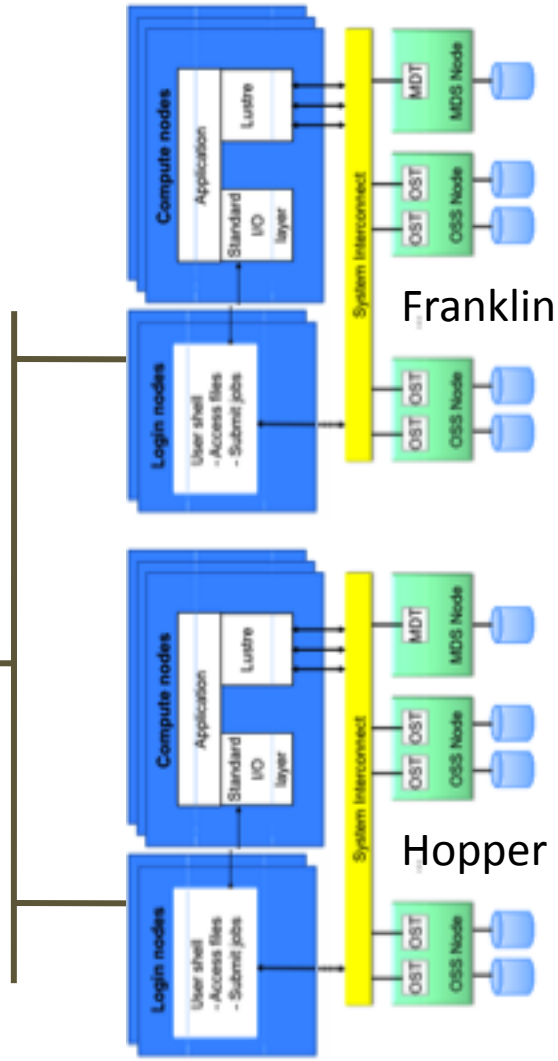
- "Local" file systems
  - Only one system can access
  - "Usually" highest performance

- Global file systems

No local disk

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)!



U.S. DEPARTMENT OF ENERGY | Office of Science

** 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

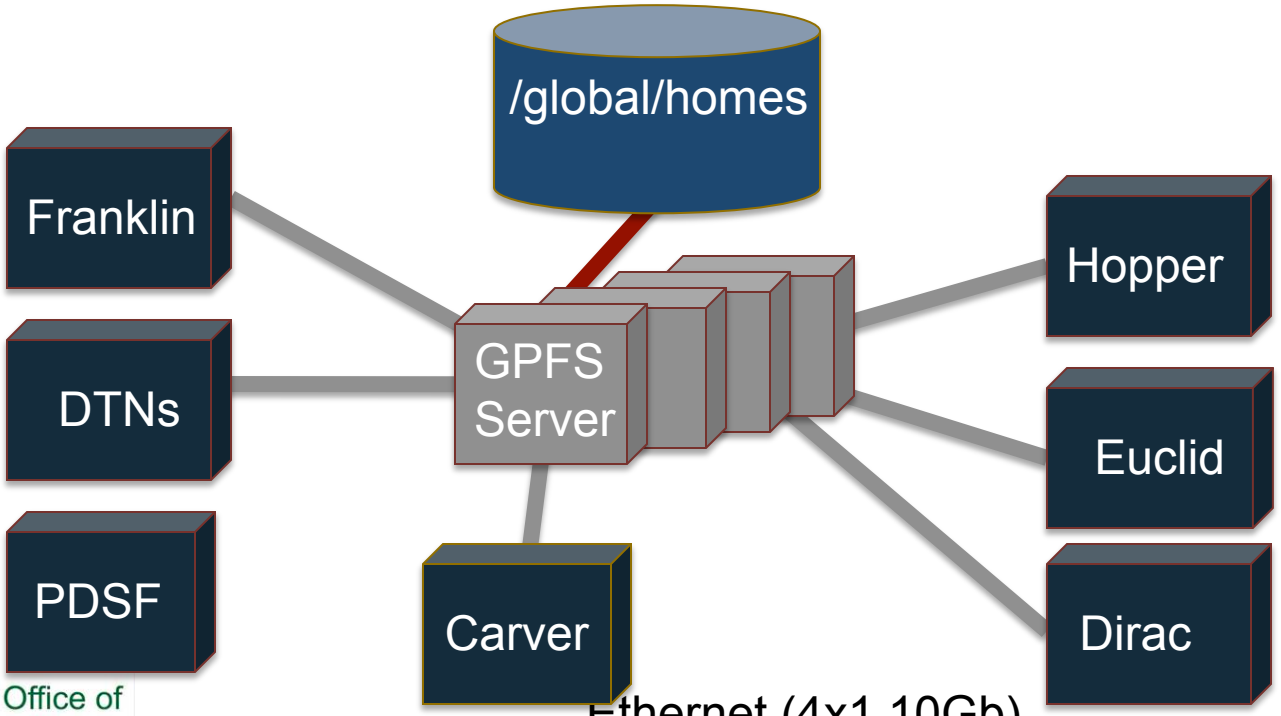
- /global/scratch

- /project

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

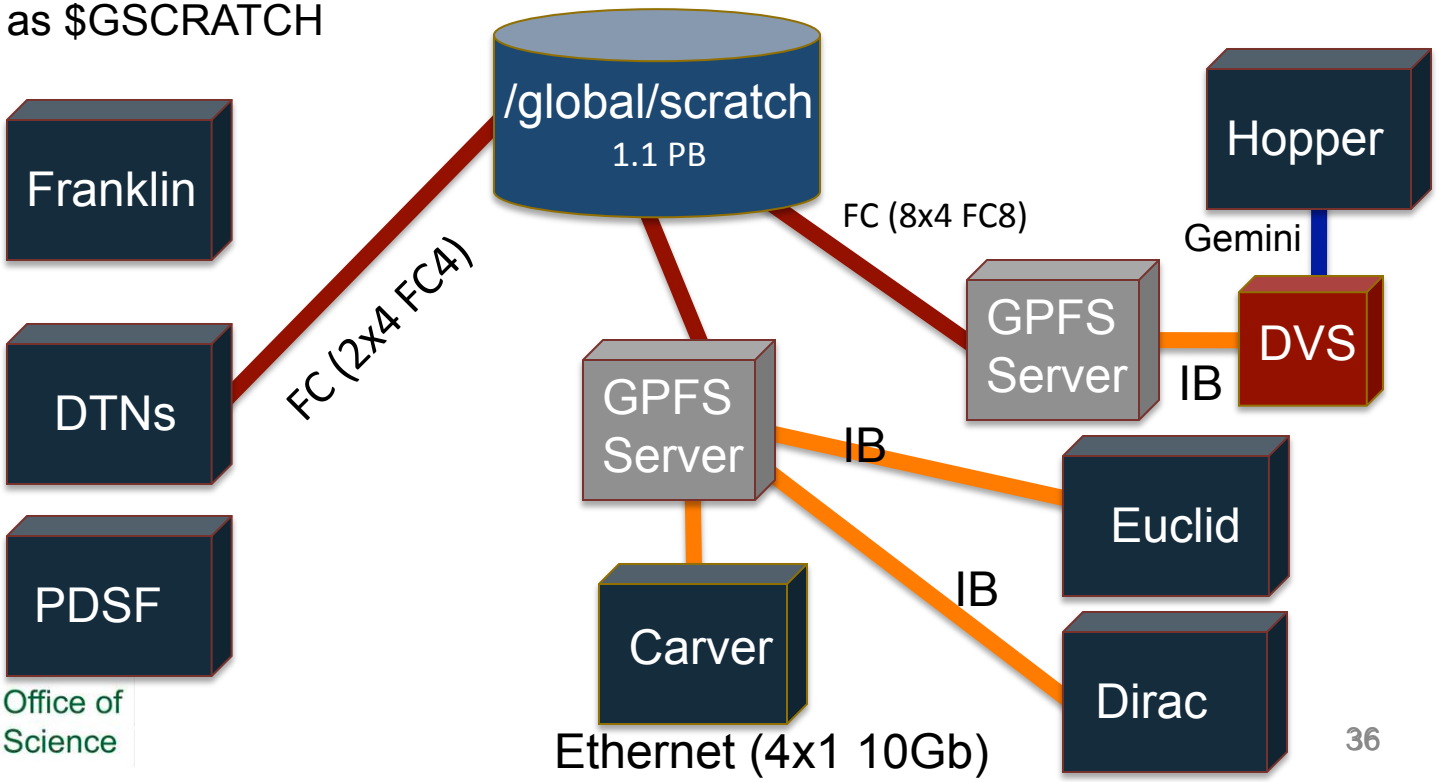
NGF Global Homes

- **/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.



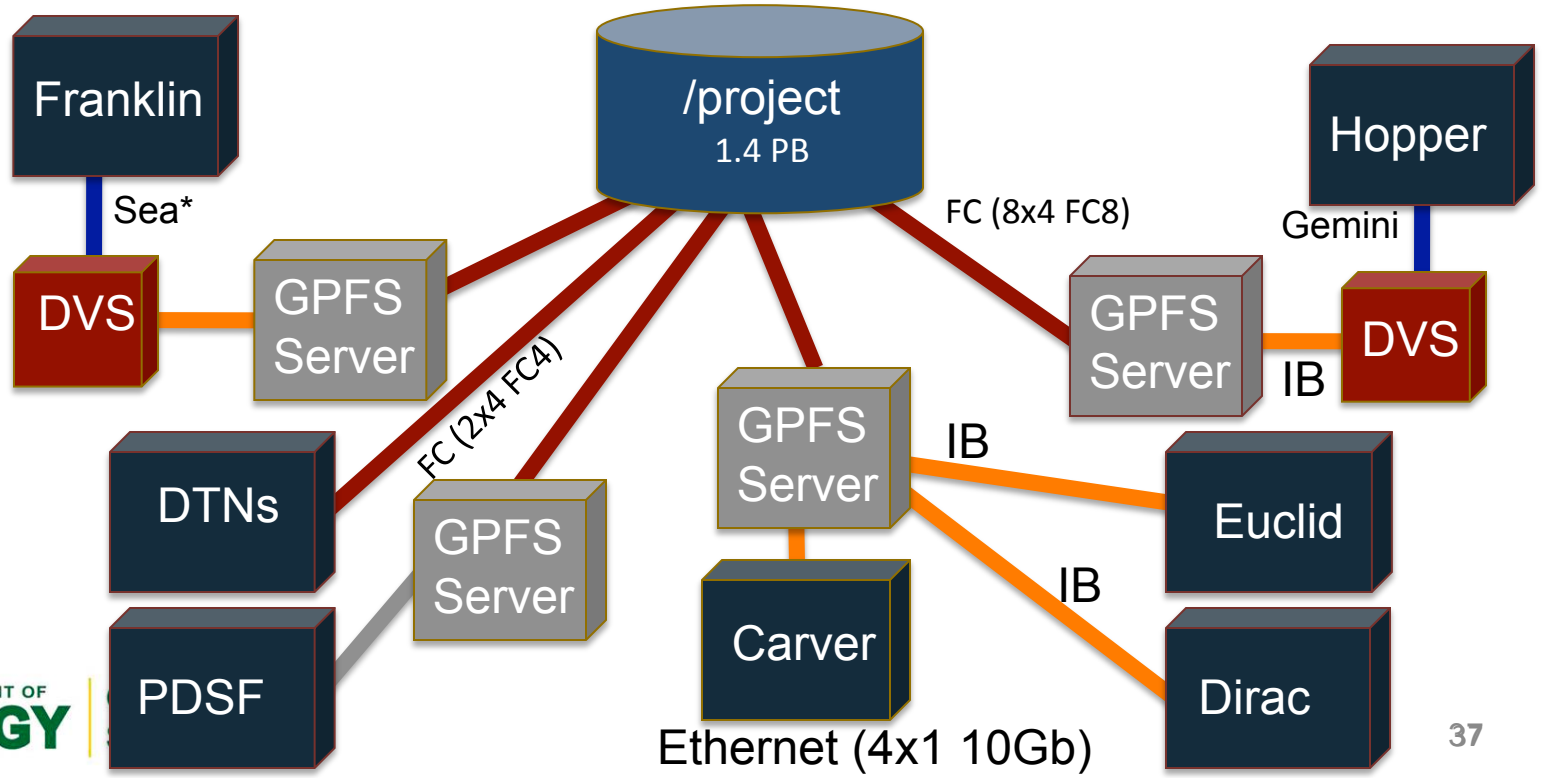
NGF Global Scratch

- **/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



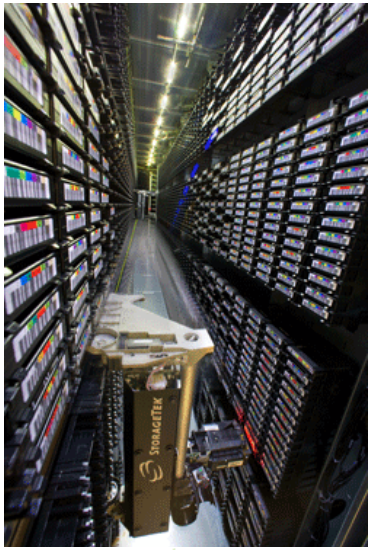
NGF Project

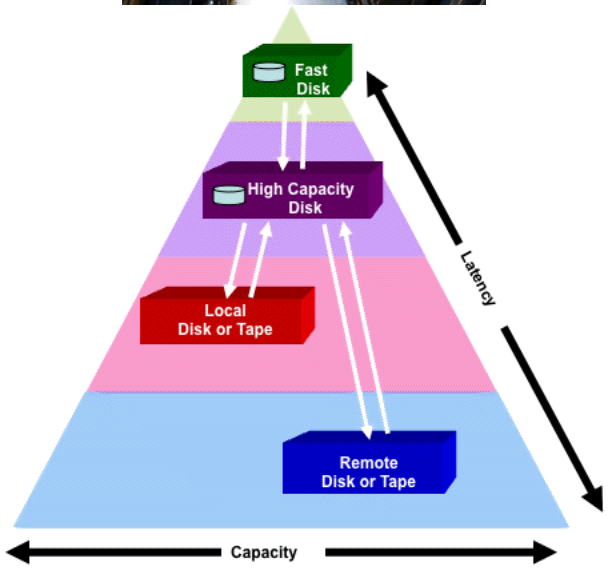
- **/project: NERSC-wide sharing and long-term data storage**
- Obtain via special request for sharing data between platforms, users, or outside
- Not purged, quota enforced (4TB default per project), backed up daily
- Serves 200 projects; 1.4 PB (+2.8!!) total capacity; ~5 TB average daily IO



Archival Storage: HPSS

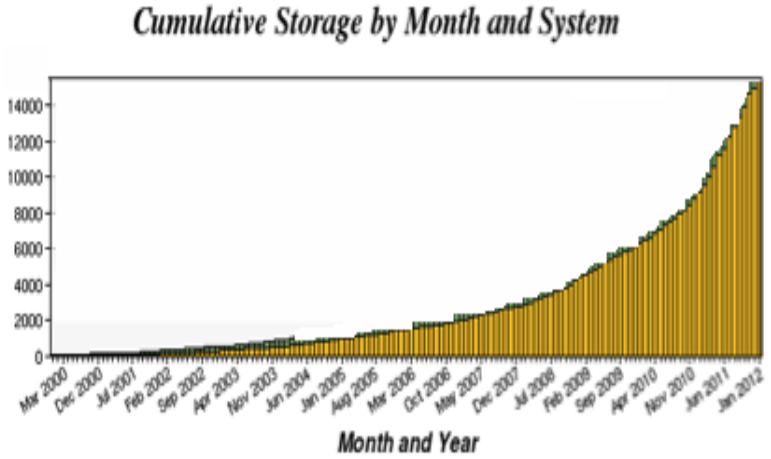
- For permanent, archival storage
- Uses magnetic tape, disk with 150TB fast-access disk cache
  - ~15 PB data in 140 M files
  - Increases at ~1.7X per year
  - Average data xfer rate: 100 MB/sec
- Cartridges are loaded unloaded into tape drives by sophisticated robotics
- Use HPSS to back up your code,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



U.S. DEPARTMENT OF ENERGY | Office of Science

* clients available for download

** not ssh

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

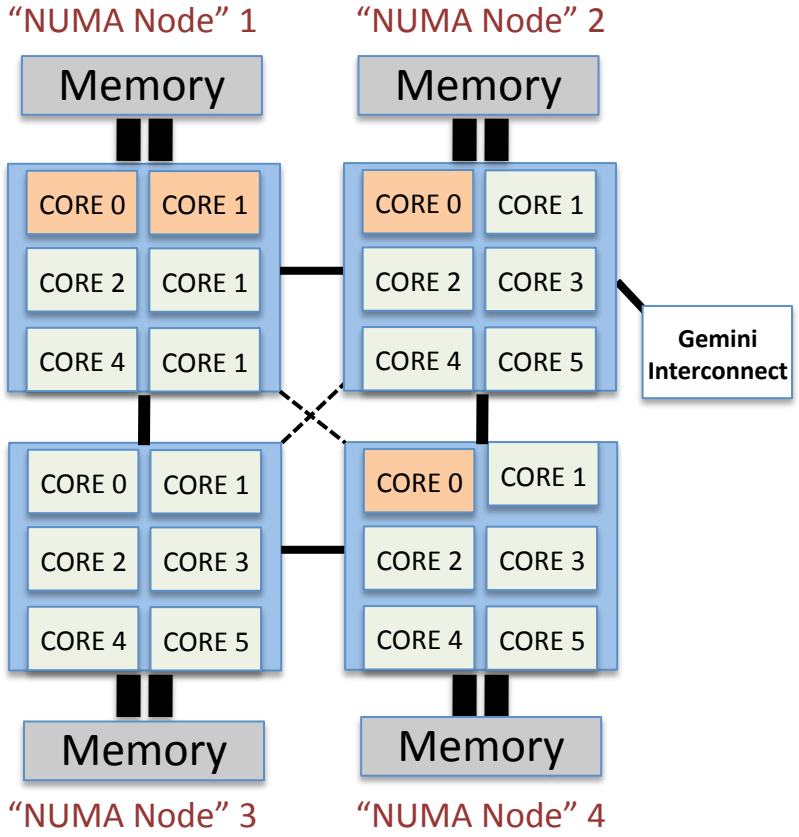
File System	Home	Local Scratch	Global Scratch	Project
Scope	Global	Local	Global	Global
Default Quota	40GB 1M inodes	5TB 5M inodes	20TB 2M inodes	4TB 4M inodes
Intended Purpose	• dot files • source codes • compiling • input files	• batch jobs • I/O intensive • temporary storage of large files	•batch jobs •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

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

- **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.