JGI Training Series

Jay Srinivasan

Computational Systems Group
Lawrence Berkeley National Lab

24 February 2012

U.S. DEPARTMENT OF ENERGY | Office of Science

JGI DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

NERSC

BERKELEY LAB

Until all users are migrated to NERSC we plan to hold weekly Friday sessions

1:30-1:45 - Intro and presentation of the transition schedule

1:45 - 2:15 - Best practices when using the shared batch system

2:15 - 2:30 - Job arrays, use, control and monitoring of jobs

2:30 - 2:45 - Review of queue policies (long jobs, highmem jobs)

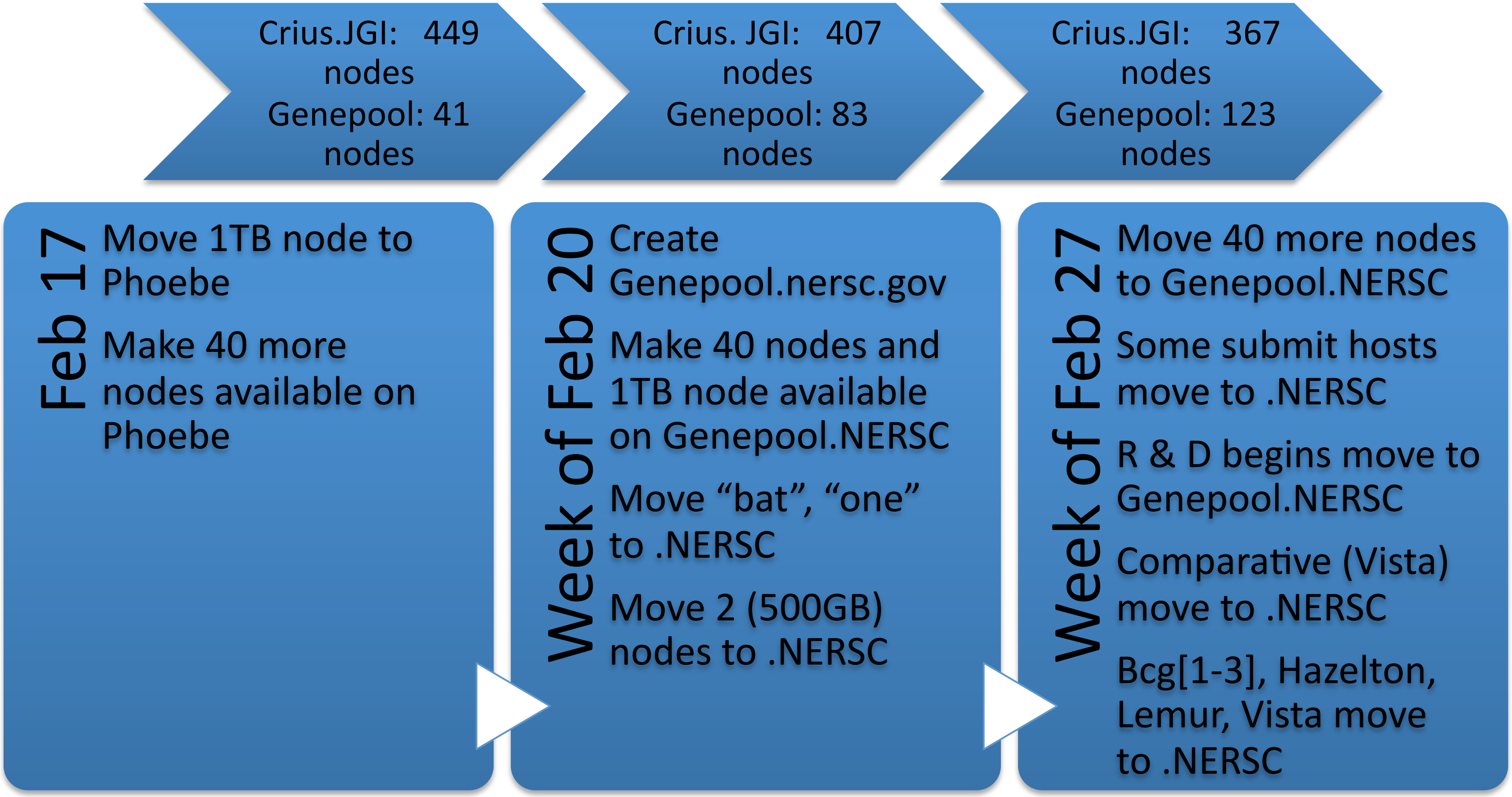
2:45 - 3:00 - Data transfer from DTN nodes including moving data off NetApps (/home) to /house or /projectb

3:00 - 5:00 - Open question hours, drop by, ask questions, trouble shoot problems with NERSC and IT staff.

Transition Schedule

- **We will perform the move of Crius workload to NERSC in stages, one week at a time.**
  - Each week, we will move part of Crius (JGI) to Genepool (NERSC). This will be done in units of ½ rack or 1 whole rack (40 or 80 nodes).
  - During that week, we will also have 1-2 JGI groups/pipelines start running on the resources at NERSC
    - NERSC staff will work closely with the group to ensure all problems are ironed out
    - NERSC staff will also begin preparations for the next group/ pipeline to move
  - During each move period, we will also move a fraction of the large-memory nodes to the NERSC space
  - During each move we will also move specific submit hosts associated with the groups/pipelines.
    - In some cases move of submit hosts/analysis machines can be done beforehand

Move Details



Crius.JGI: 449 nodes
Genepool: 41 nodes

Crius. JGI: 407 nodes
Genepool: 83 nodes

Crius.JGI: 367 nodes
Genepool: 123 nodes

Feb 17

- Move 1TB node to Phoebe
- Make 40 more nodes available on Phoebe

Week of Feb 20

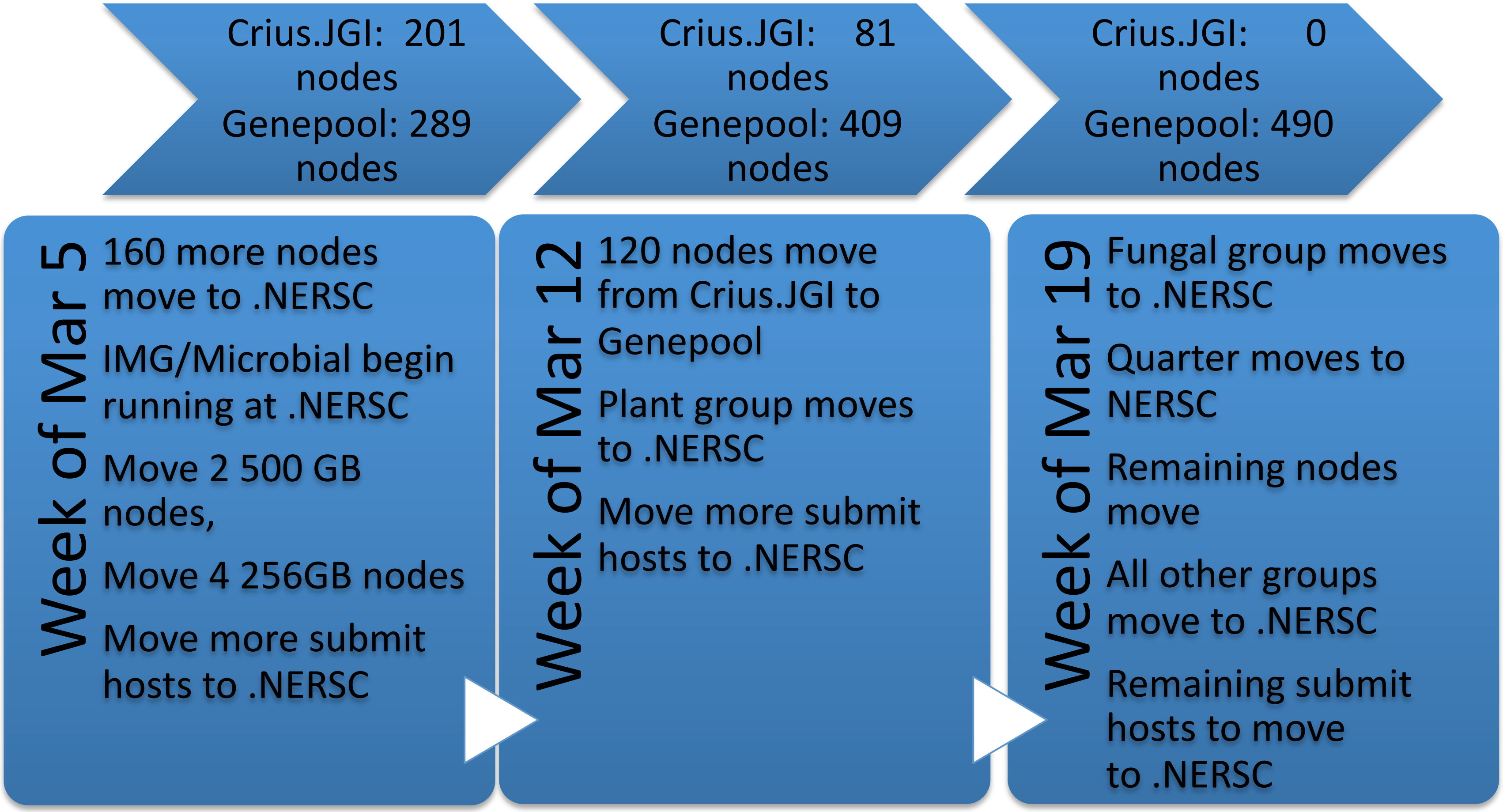
- Create Genepool.nersc.gov
- Make 40 nodes and 1TB node available on Genepool.NERSC
- Move “bat”, “one” to .NERSC
- Move 2 (500GB) nodes to .NERSC

Week of Feb 27

- Move 40 more nodes to Genepool.NERSC
- Some submit hosts move to .NERSC
- R & D begins move to Genepool.NERSC
- Comparative (Vista) move to .NERSC
- Bcg[1-3], Hazelton, Lemur, Vista move to .NERSC

All Groups are encouraged to continue running on Phoebe/Genepool.NERSC

Move Details



Crius.JGI: 201 nodes
Genepool: 289 nodes

Crius.JGI: 81 nodes
Genepool: 409 nodes

Crius.JGI: 0 nodes
Genepool: 490 nodes

Week of Mar 5

- 160 more nodes move to .NERSC
- IMG/Microbial begin running at .NERSC
- Move 2 500 GB nodes,
- Move 4 256GB nodes
- Move more submit hosts to .NERSC

Week of Mar 12

- 120 nodes move from Crius.JGI to Genepool
- Plant group moves to .NERSC
- Move more submit hosts to .NERSC

Week of Mar 19

- Fungal group moves to .NERSC
- Quarter moves to NERSC
- Remaining nodes move
- All other groups move to .NERSC
- Remaining submit hosts to move to .NERSC

All Groups are encouraged to continue running on Phoebe/Genepool

Best Practices for Batch system

- The batch system controls a shared resource. You should be aware of the effect you have on other users of the system.
- Qstat and Qmod are best run interactively. Don't run them in tight loops via scripts. Try to use cached data as much as possible (we are working to provide a consistent query method for all groups)
- Don't automatically clear errored jobs without checking for why the error occurred.
- If possible checkpoint your pipelines.
- Limit coredumps on your production workload. If you get an error, you can enable coredumps and rerun in a `debug` mode.
- Use Job arrays as much as possible.
- Very short jobs (less than 1 min) put a considerable load on the scheduler. Consider combining them into longer jobs.

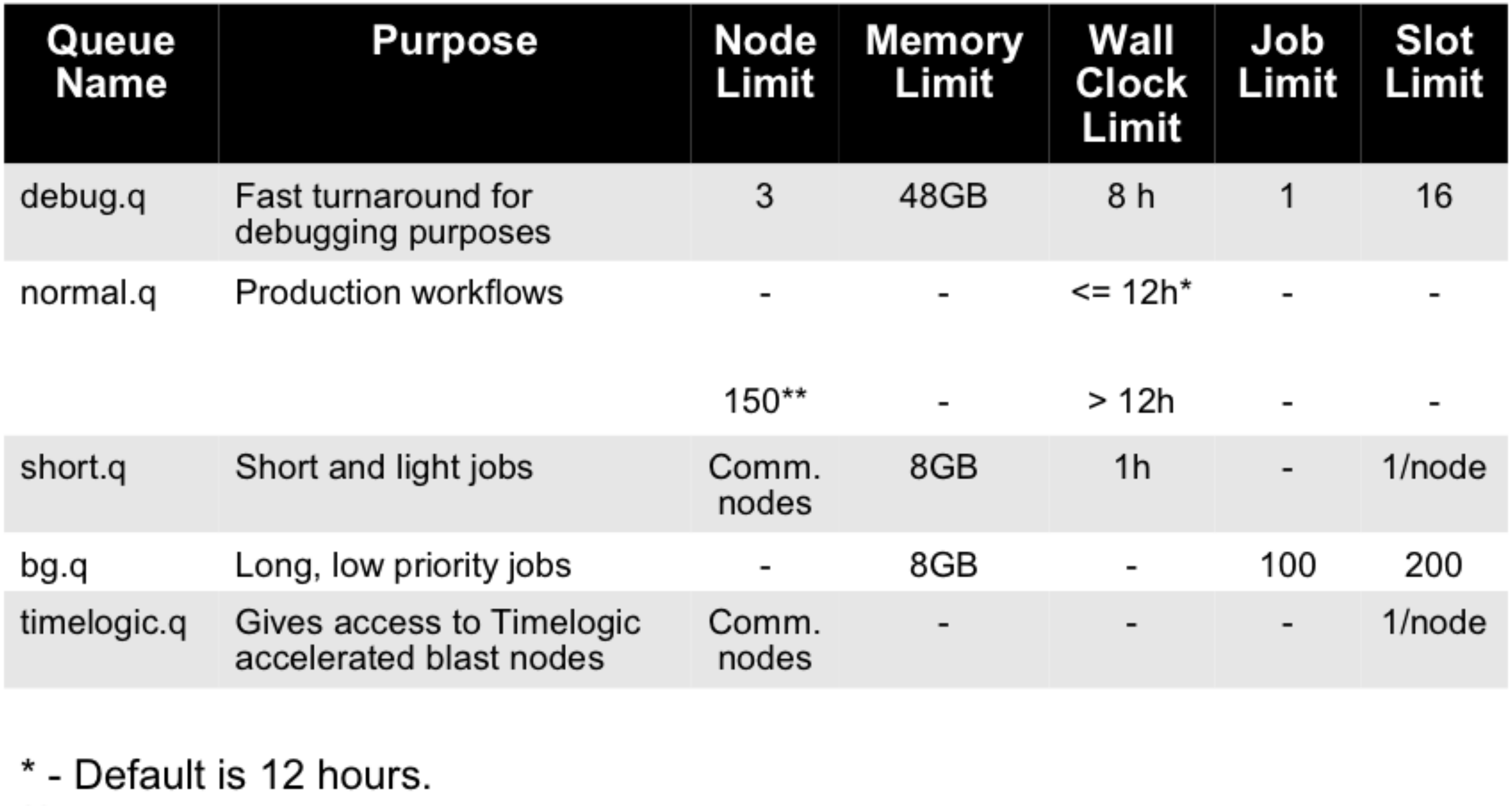
Best Practices for Batch system

- Try to get a good estimate of what resources your jobs need (memory, runtime).
- Set both soft and hard limits, so you get warned when you are close to reaching resource limits. Trap the warnings and perform necessary action to gracefully exit.
- Use “-w e” option to qsub. This performs checking and prevents unschedulable jobs from being submitted.
- Use job dependencies to order your jobs and ensure pipeline requirements are met.

Job control

- Submitting a very large number of jobs is hard on the scheduler. Use of Job Arrays helps reduce the load on the scheduler.
- Use $TASK_ID/$SGE_TASK_ID to distinguish between the subtasks of the jobs.
- Tasks in a job array are scheduled individually.
- Use job dependencies to order your jobs and ensure pipeline requirements are met.

Queuing Policies



Queue Name	Purpose	Node Limit	Memory Limit	Wall Clock Limit	Job Limit	Slot Limit
debug.q	Fast turnaround for debugging purposes	3	48GB	8 h	1	16
normal.q	Production workflows	-	-	<= 12h*	-	-
		150**	-	> 12h	-	-
short.q	Short and light jobs	Comm. nodes	8GB	1h	-	1/node
bg.q	Long, low priority jobs	-	8GB	-	100	200
timelogic.q	Gives access to Timelogic accelerated blast nodes	Comm. nodes	-	-	-	1/node

* - Default is 12 hours.

Jobs running longer than 12 hours or requesting large amounts of memory could see longer wait times

Queue Policies

- **We encourage use of h_vmem as a resource specification. It is related to ram.c (which is not enforced). We plan to deprecate the use of ram.c.**
  - h_vmem is a per node specification – you just need to specify how much your job uses regardless of number of slots used
  - ram.c is a per slot specification – we take care of multiplication for parallel jobs
- **Long running jobs are currently sequestered to 150 nodes. We plan to give each group its own allocation of nodes which can run long running jobs. The number of these will be proportional to the number of nodes brought into the merged cluster**

NERSC has set up 2 fast “data transfer nodes” just for JGI users

Login to dtn03.nersc.gov or dtn04.nersc.gov

Type >df to see all the mounted file systems

Back up data to HPSS (you authenticated at last week’s training don’t remember? Type hsi and then enter your NIM password)

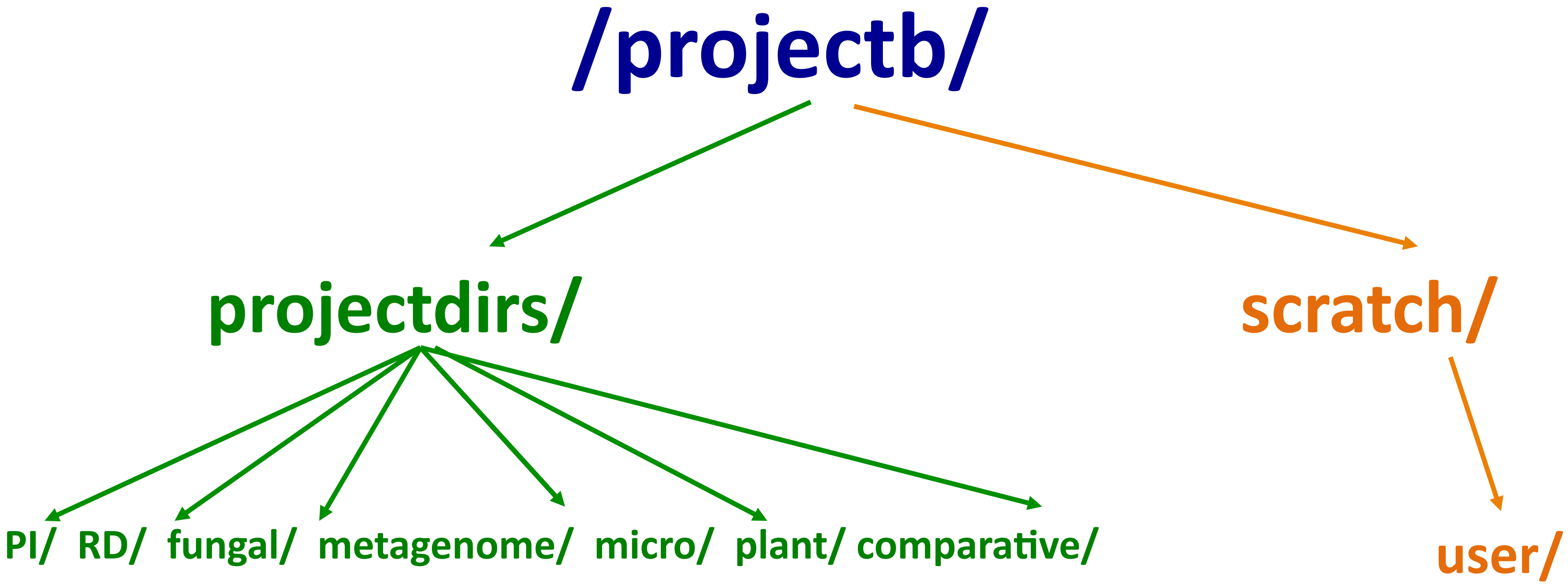
```
> cd /house/path/to/your/data
> hsi put <filename>
```

Or archive an entire directory

```
> htar –cvf  tarname.tar directory/
```

There are two areas of storage within the “project” layout of the “projectb” file system

ssh phoebe.nersc.gov



/projectb/

projectdirs/

scratch/

PI/ RD/ fungal/ metagenome/ micro/ plant/ comparative/

user/

- Group directories
- Not purged
- Subject to quota

- User directories
- cd $SCRATCH
- Purged, 12 weeks
- 1 TB, 500,000 inode quota

Request a projectb directory for your group through the Jira ticket system

Request a larger /scratch quota through the Jira ticket system

Use the fast data transfer nodes to move data between file systems

Login to dtn03.nersc.gov or dtn04.nersc.gov

Type >df to see all the mounted file systems

You can move data to 3 file systems $HOME “project” “scratch”

```
> mv /old/path/filename  /new/path/filename
```

Data Transfer

- **For data movement between JGI and NERSC or between different storage systems, use the Data Transfer Nodes**
    - dtn03.nersc.gov & dtn04.nersc.gov
    - Have NetApps mounted (/JGI/home, /JGI/psf, /JGI/storage)
    - Have Isilon mounted (/house, /ifs)
    - Have NGF mounted (/global/homes, /project, /projectb, /global/scratch)
    - HPSS access from here (hsi and htar)
- **NetApps will be retired by April 30, 2012.**
    - No support contract for hardware that is seven years old
    - Users should move data, if needed, from NetApps to one of the other file systems or HPSS
- **Move data from /house to HPSS**
    - SDM and their SRF files (370 TB), seeing 1.2GB/sec