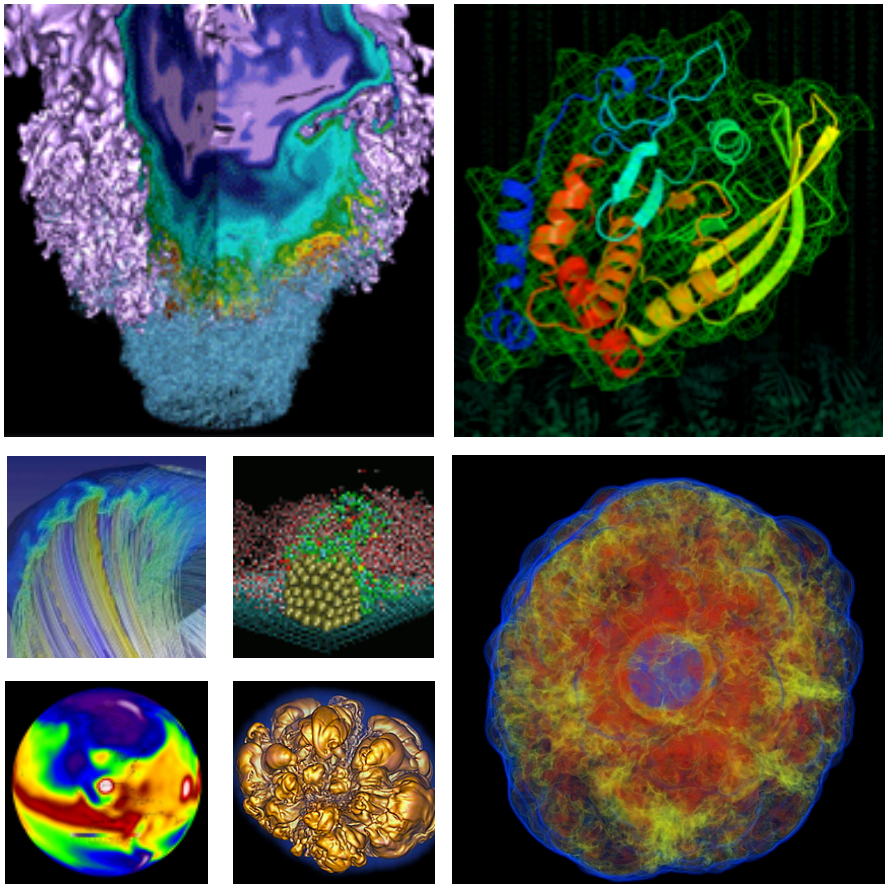
Data Management at NERSC



Lisa Gerhardt

NERSC User Services Group

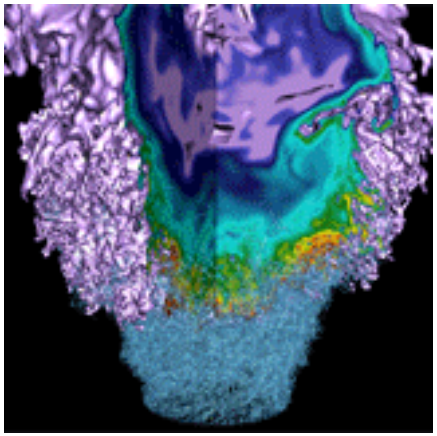
New User Training

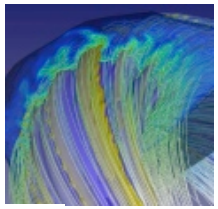
July 15, 2014

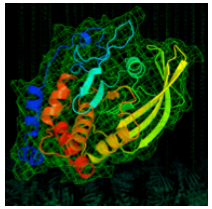
Where Do I Put My Data?

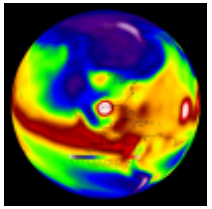
- **Overview of NERSC file systems**
  - Local vs. Global
  - Permanent vs. Purged
- **HPSS Archive System**
  - What is it and how to use it
- **Data Sharing**

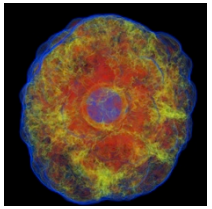
NERSC File Systems

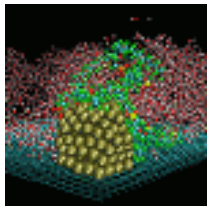




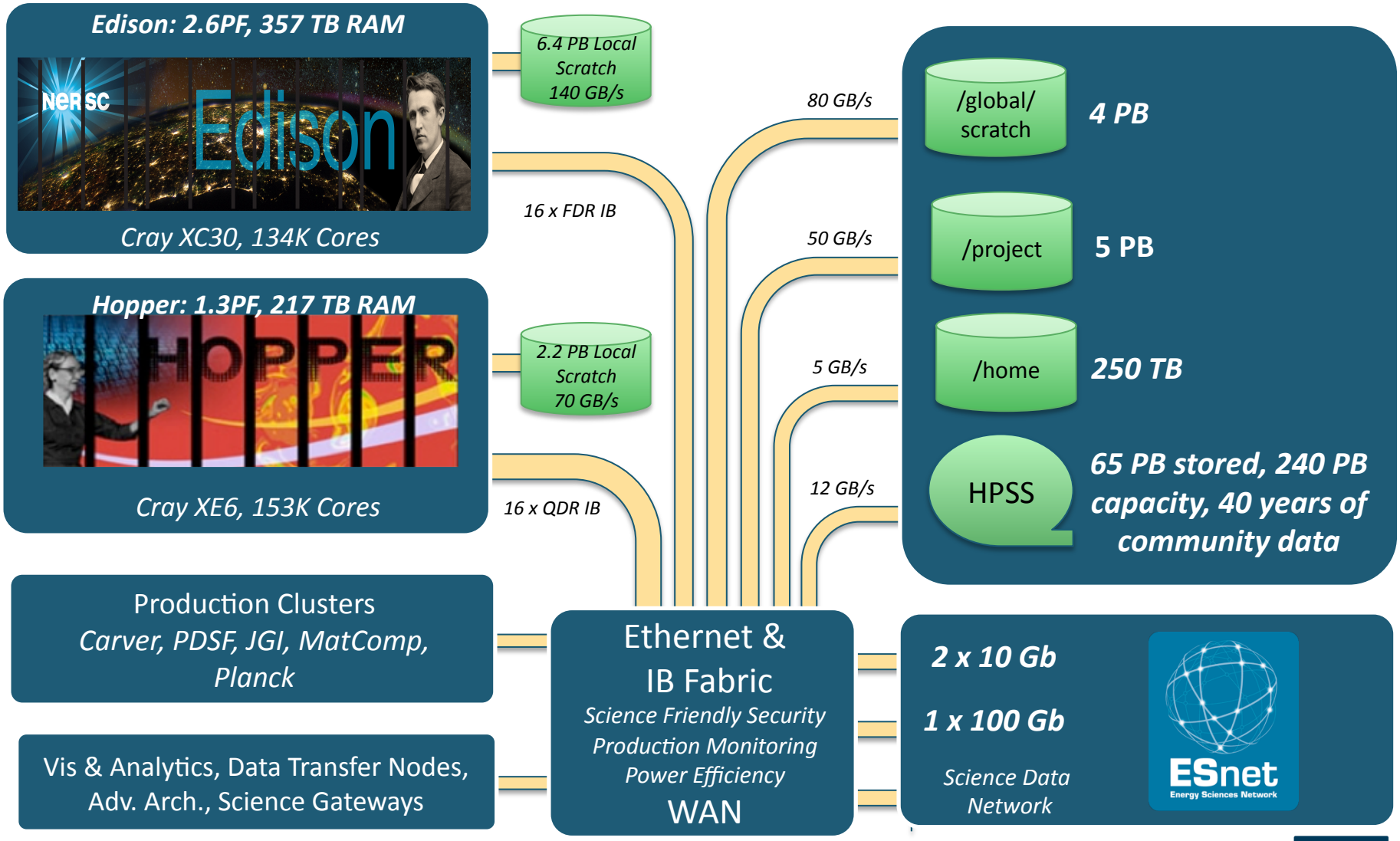








The compute and storage systems 2014



Protect Your Data!

- Some file systems are backed up
- Some file systems are not backed up
- Restoration of individual files/directories may *not* be possible
- Hardware failures and human errors *will* happen

BACK UP YOUR FILES TO HPSS!

Global File Systems

- **NERSC Global Filesystem (NGF)**
  - Based on IBM's General Parallel File System (GPFS)
- **Accessible on all NERSC systems**
- **Provides directories for home, global scratch, and project**
- **Shared by ~5000 active NERSC users**

Global Homes File System Overview

- **Provided by two ~100 TB file systems**
  - 5 GB/s aggregate bandwidth
- **Access with $HOME, ~/<file_in_home_dir>**
- **Other name**

```
/global/homes/d/dpturner
```

- **Low-level name**

```
/global/u1/d/dpturner
/global/u2/d/dpturner ->
  /global/u1/d/dpturner
```

Global Homes Use

- **Shared across all platforms**
  - Dot files that control user environment
  - **`$HOME/edison`**, **`$HOME/hopper`**, etc.
- **Tuned for small file access**
  - Compiling/linking
  - Configuration files
  - Do not send batch job output to $HOME!

Global Homes Policies

- **Quotas enforced**
  - 40 GB
  - 1,000,000 inodes (i.e. files and directories)
  - Quota increases rarely (i.e., never) granted
  - Monitor with **`myquota`** command
- **"Permanent" storage**
  - No purging
  - Backed up
  - Hardware failures and human errors *will* happen

BACK UP YOUR FILES TO HPSS!

Project File System Overview

- **Provides 5.1 PB high-performance disk**
  - 50 GB/s aggregate bandwidth
- **Available on all NERSC systems**
- **Intended for sharing data between platforms, users, or with the outside world**
- **Beginning this year every MPP repo gets a project directory**

  `/project/projectdirs/m9999`

Project Use

- **Tuned for large streaming file access**
  - Sharing data within a project or externally
  - Running I/O intensive batch jobs
  - Data analysis/visualization
- **Access controlled by Unix file groups**
  - Group name usually same as directory
  - Requires administrator (usually the PI or PI Proxy)
  - Can also use access control list (ACL)

Project Policies

- **Quotas enforced**
  - 1 TB
  - 1,000,000 inodes
  - Quota increases may be requested
  - Monitor with **`prjquota`** command

    ```
    % prjquota bigsci
    ```
- ***Permanent* storage**
  - No purging
  - Backed up if quota <= 5 TB
  - Hardware failures and human errors *will* happen

BACK UP YOUR FILES TO HPSS!

Global Scratch File System Overview

- **Provides 4 PB high-performance disk**
  - 80 GB/s aggregate bandwidth
- **Access with `$GSCRATCH`**
- **Low-level name**

  `/global/scratch2/sd/dpturner`

Global Scratch Use

- **Shared across all systems**
  - Primary scratch file system for Carver
- **Tuned for large streaming file access**
  - Running IO intensive batch jobs
  - Data analysis/visualization

Global Scratch Policies

- **Quotas enforced**
  - 20 TB
  - 4,000,000 inodes
  - Quota increases may be requested
  - Monitor with **`myquota`** command
- ***Temporary* storage**
  - Bi-weekly purges of *all* files that have not been accessed in over 12 weeks
    - List of purged files in $GSCRATCH/purged.<timestamp>
  - Hardware failures and human errors *will* happen

BACK UP YOUR FILES TO HPSS!

Local File Systems on Cray Systems

- **Edison and Hopper have local scratch**
- **Edison has two *scratch* file systems**
  - Users randomly assigned
  - Each has 2.1 PB (1 PB on Hopper)
  - Each has 48 GB/s aggregate bandwidth (35 GB/s Hopper)
- **Edison has extra high-performance scratch (scratch3)**
  - 3.2 PB, 72 GB/s aggregate bandwidth
- **Provided by Cray, based on Lustre**
- **Generally, IO access for batch jobs on Hopper and Edison will be fastest for local scratch**

Edison Scratch Use

- **Each user gets a scratch directory in /scratch1 *or* /scratch2 (Hopper: /scratch or /scratch2)**

  `/scratch2/scratchdirs/dpturner`

  – **Best name: `$SCRATCH`**
- **Access to /scratch3 must be requested**
  – Large datasets
  – High bandwidth
- **Tuned for large streaming file access**
  – Running I/O intensive batch jobs
  – Data analysis/visualization

Edison Scratch Policies

- **Quotas enforced in $SCRATCH by submit filter**
  - 10 TB (5 TB Hopper)
  - 10,000,000 inodes (5M inodes Hopper)
  - Quota increases may be requested
  - Monitor with **`myquota`** command
  - No quota enforcement in /scratch3
- ***Temporary* storage**
  - Daily purges of *all* files that have not been accessed in over 12 weeks
    - List of purged files in $SCRATCH/purged.<timestamp>
  - Hardware failures and human errors *will* happen

BACK UP YOUR FILES TO HPSS!

Long-Term File Systems

- **Global home directories ($HOME)**
  - Source/object/executable files, batch scripts, input files, configuration files, batch job summaries (*not* for running jobs)
  - Backed up
  - 40 GB permanent quota
- **Global project directories**
  - Sharing data between people and/or systems, short term data storage
  - Backed up if quota less than or equal to 5 TB
  - All MPP repos have one, 1 TB default quota

Short-Term File Systems

- **Local scratch directories**
  - Cray (Edison, Hopper) only
  - Large, high-performance parallel Lustre file system
  - Not backed up; files purged after 12 weeks
  - Hopper: 5 TB default quota; Edison: 10 TB default quota
  - $SCRATCH, $SCRATCH2
- **Global scratch directories**
  - All systems
  - Large, high-performance parallel GPFS file system
  - Not backed up; files purged after 12 weeks
  - 20 TB default quota
  - $GSCRATCH

Where Do I Put My Data?

Local Scratch
Fastest IO
Only visible on one machine
Purged

Project
Medium IO
Visible on all machines
Never purged
External sharing

Global Scratch
Fast IO
Visible on all machines
Purged

Home
Source code, config. files
No batch output

File Systems Summary

File System	Path	Type	Default Quota	Backups	Purge Policy
Global Homes	$HOME	GPFS	40 GB / 1M inodes	Yes	Not purged
Global Scratch	$GSCRATCH	GPFS	20 TB / 4M inodes	No	12 weeks from last access
Global Project	/project/ projectdirs/ *projectname*	GPFS	1 TB / 1M inodes	Yes, if quota less than or equal to 5TB	Not purged
Hopper Scratch	$SCRATCH and $SCRATCH2	Lustre	5 TB / 5M inodes (combined)	No	12 weeks from last access
Edison Scratch	$SCRATCH	Lustre	10 TB / 5M inodes (none in /scratch3)	No	12 weeks from last access

Resources

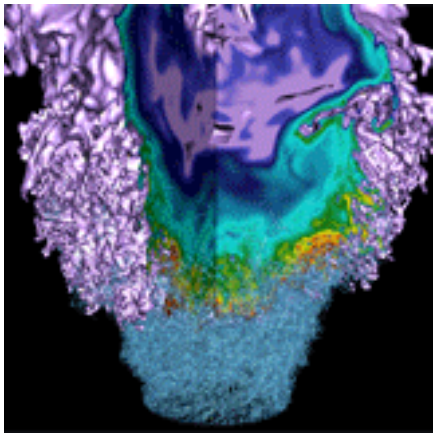
http://www.nersc.gov/users/data-and-file-systems/

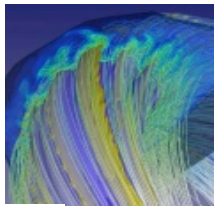
http://www.nersc.gov/users/data-and-file-systems/file-systems/

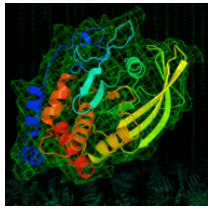
http://www.nersc.gov/users/computational-systems/edison/file-storage-and-i-o/

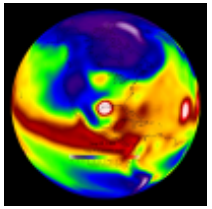
http://www.nersc.gov/users/computational-systems/hopper/file-storage-and-i-o/

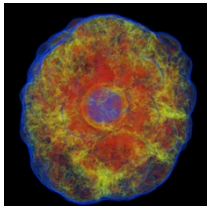
HPSS: The NERSC Archive System

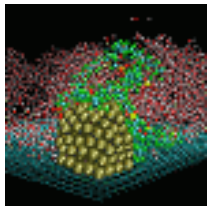






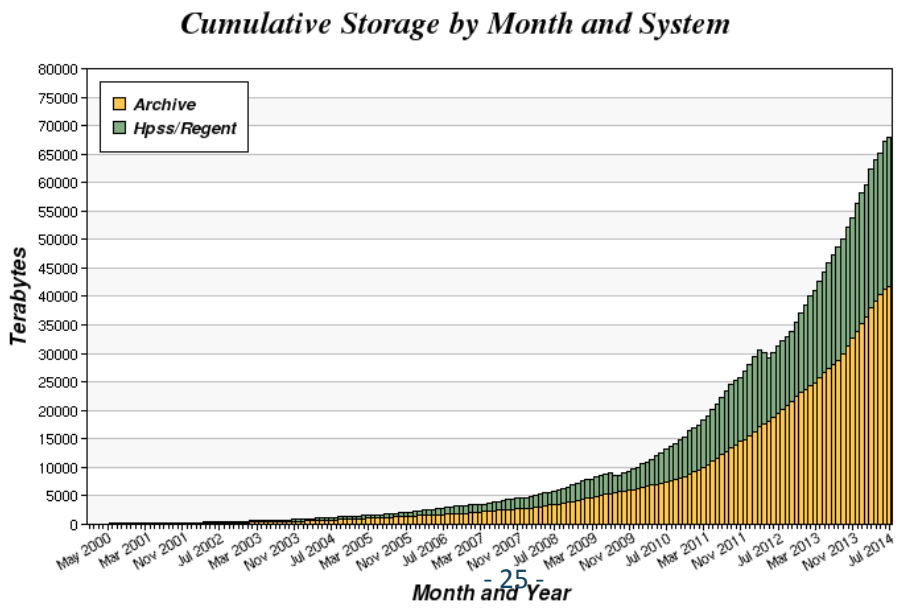






Archiving Data is Necessary

- **Data growth is exponential and file system space is finite**
  - 80% of stored data is never accessed after 90 days
  - Cost of storing infrequently accessed data on flash or spinning disk is prohibitive
  - Store important data in an archive to free faster resources for processing workload
    - Data from publications, unique experimental, or simulation data
- **NERSC provides the HPSS archive system for data archiving**

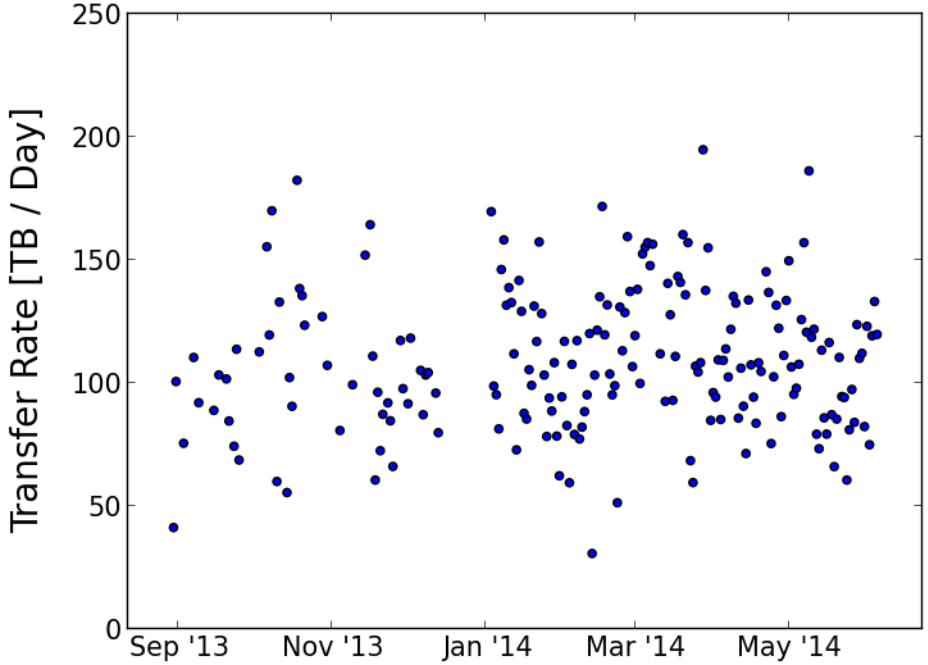


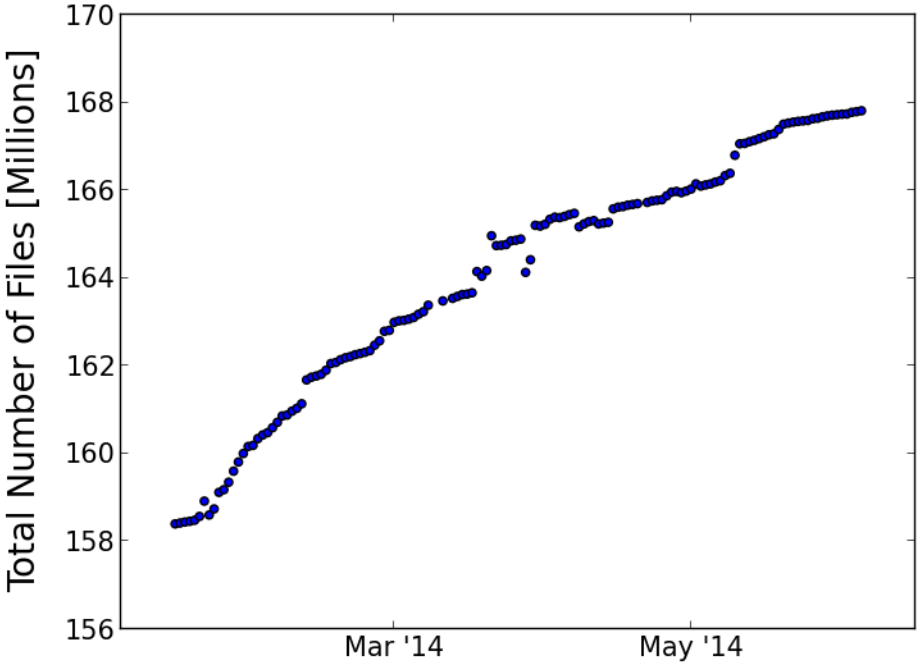
40 PB of data
Started in 1998, but oldest file is from the 70s

Features of the NERSC archive

- **NERSC implements an online or "active archive"**
  - Parallel high-speed transfer and fast data access
    - Data is transferred over parallel connections to the NERSC internal 10Gb network
    - Access to first byte in seconds or minutes as opposed to hours or days
  - Tiered internal storage facilitates high speed data access:
    - Initial data ingest to high-performance disk cache
    - Data migrated to automated enterprise tape system and managed by HSM software (HPSS) based on file age and usage
  - Indefinite data retention policy
- **The archive is accessible to all NERSC users**
- **Often referred to as HPSS**

HPSS is Heavily Used





Accessing HPSS from NERSC Systems

- **HSI**
  - Fast, parallel transfers, unix-like interface
  - Store from file system to archive:

    -bash-3.2$ **hsi**

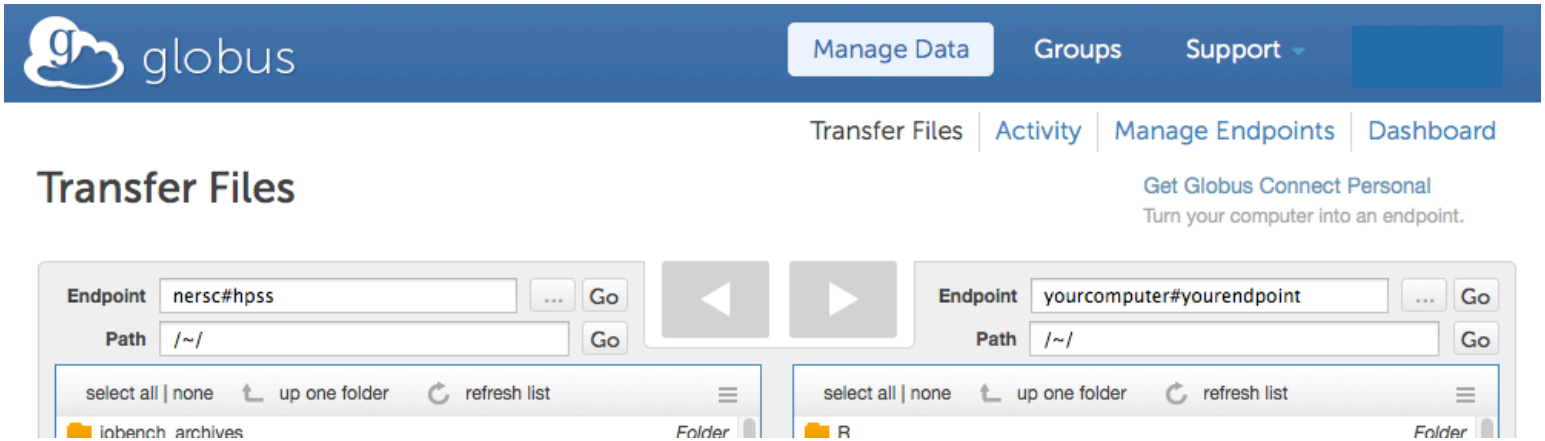
    A:/home/n/nickb-> **put myfile**

    put 'myfile' : '/home/n/nickb/myfile' (2097152 bytes, 31445.8 KBS (cos=4))
- **HTAR**
  - Parallel, puts tar file directly into HPSS, excellent for groups of small files
  - Syntax: *htar [options] <archive file> <local file|dir>*

    -bash-3.2$ **htar –cvf /home/n/nickb/mydir.tar ./mydir**

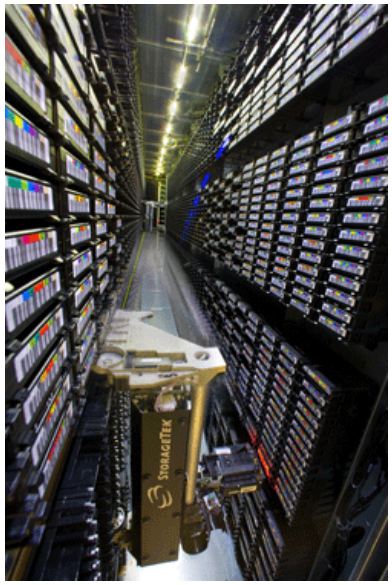
Accessing HPSS from Outside NERSC

- **HSI and HTAR precompiled binaries available for most systems**
- **ftp:** non-parallel, but common
- **gridFTP:** parallel, requires credential
- **Globus:** parallel, requires endpoint



Tape IO Characteristics

- **Ultimately all data in HPSS is written to tape**
- **Tape is linear media**
  - Behaves differently than disk:
    - Data cannot be re-written in place, it is appended at the end
    - Reading and writing are sequential operations – no random access
- **Tape drives behave differently than disk drives**
  - Take time to seek to file locations on tape
  - Take time to ramp up to full speed
  - Tape drives stop after reading or writing each file
- **HPSS will not respond like a normal file system**
  - Presents itself as one, but some things can have unexpected results



Size Matters

- **Sweet Spot**
  - Tape drives perform best when operating at full rate for long durations
  - Large file are best for tape drive performance
  - Many small files causes frequent stops and low-speed operations, can take a very long time to retrieve
  - File bundles in the **100s of GB** currently provide best performance
- **Group small files for optimal storage**
  - Use HTAR, GNU tar, or zip to bundle groups of small files together to optimize tape and network performance
- **There is such a thing as too big**
  - Files spanning multiple tapes incur tape mount delays

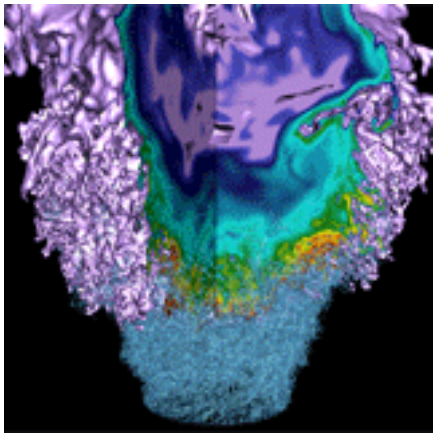
Best Practices

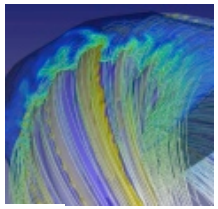
- **Group small files together and avoid excessive writes**
  - Use htar or tar to group into ~100s of GB
- **Order your retrievals**
  - Grab files from a tape in order of tape position
  - Grab all files from a tape while tape is mounted
- **Avoid excessive transfer failures**
  - Globus with unreliable network will retry many times
  - Directory permission issues
- **No exclusive access to the archive**
  - No batch system
  - Inefficient use affects performance for everyone

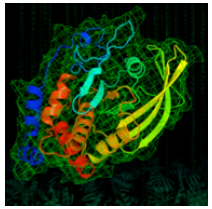
Further Reading

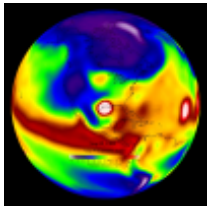
- **NERSC Website**
  - Archive documentation:
    - *http://www.nersc.gov/users/data-and-file-systems/hpss/getting-started/*
  - Data management strategy and policies:
    - *http://www.nersc.gov/users/data-and-file-systems/policies/*
  - Accessing HPSS
    - *http://www.nersc.gov/users/data-and-file-systems/hpss/getting-started/accessing-hpss/*
- **HSI and HTAR man pages are installed on NERSC compute platforms**
- **Gleicher Enterprises Online Documentation (HSI, HTAR)**
  - *http://www.mgleicher.us/index.html/hsi/*
  - *http://www.mgleicher.us/index.html/htar/*
- **"*HSI Best Practices for NERSC Users,*" LBNL Report #LBNL-4745E**
  - *http://www.nersc.gov/assets/pubs_presos/HSIBestPractices-Balthaser-Hazen-2011-06-09.pdf*

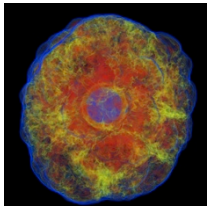
Data Sharing

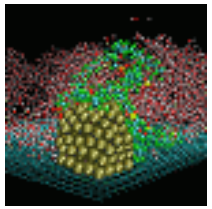












Data Sharing

- **Ensure security**
  - Do not share passwords
  - Do not share files from $HOME
- **Project directories designed for sharing**
  - Open to anyone in the repository
- **Use Unix *group* permissions**
  - Request creation of Unix group
  - Set permissions with chgrp/chmod
    - Use setgid bit

give/take

- **New, but based on *old* LLNL and LANL commands**
- **Appropriate for smaller files**

```
joe% give -u bob coolfile
```

  - File copied *to* spool location
  - Bob gets email telling him Joe has given him a file

```
bob% take -u joe coolfile
```

  - File copied *from* spool location

- **Spooled files count against *giver's* GSCRATCH quota**

Science Gateways on Project

- **Make data available to outside world**

```
mkdir /project/projectdirs/bigsci/www
chmod o+x /project/projectdirs/bigsci
chmod o+rx /project/projectdirs/bigsci/www
```

- **Access with web browser**

```
http://portal.nersc.gov/project/bigsci
```

Data Transfer

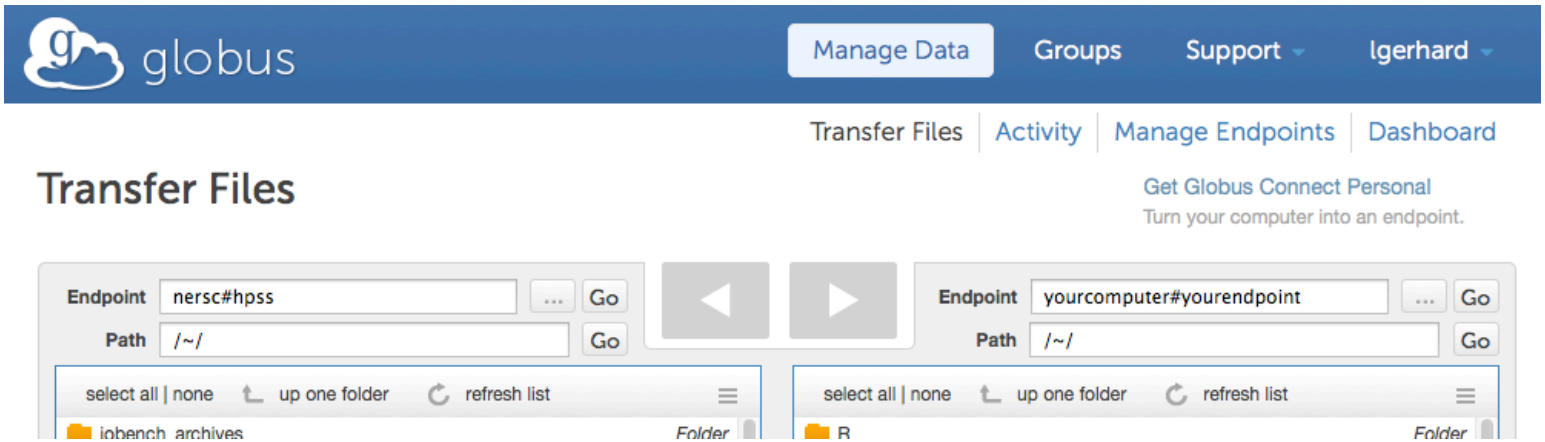
- **Global file systems**
  - Use *local* cp instead of *remote* scp
- **Use scp for small-to-medium files over short-to-medium distance**
  - Even better if HPN versions installed

```
% ssh -v
OpenSSH_5.1p1NMOD_2.9-hpn13v5, OpenSSL 0.9.8e-fips-rhel5 01 Jul
   2008
```

- **Use bbcp for larger files and/or longer distances**
  - Many tuning options
  - Complicated command line

Globus

- **Do-it-all web-based file transfer service**
- **High-performance**
  - Parallel data channels (gridftp)
- **Fire and forget model**
- **Also has a command-line interface for scripting**



Further Reading

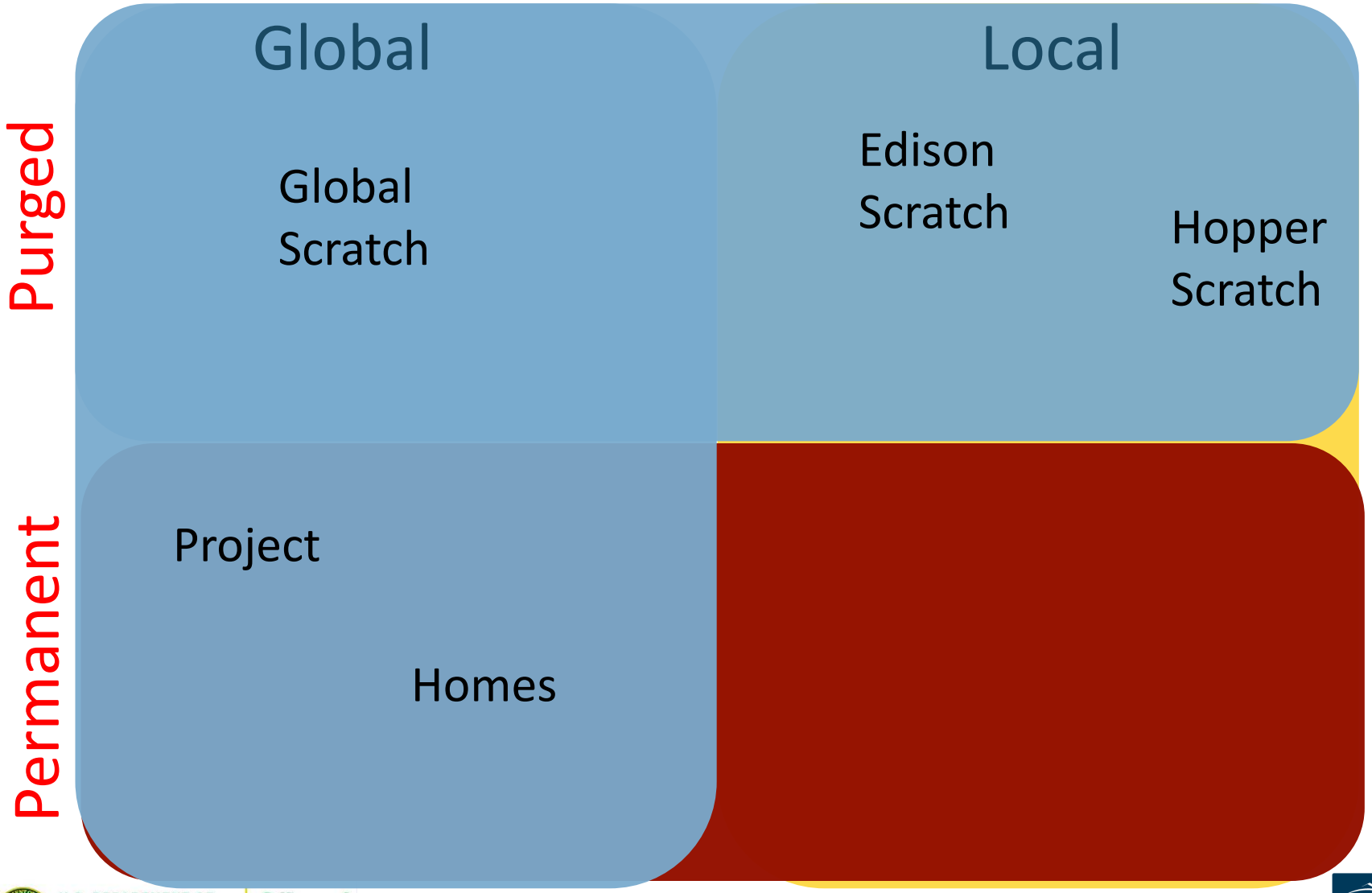
- **Sharing data**
  - https://www.nersc.gov/users/data-and-file-systems/sharing-data/
- **Transferring Data**
  - https://www.nersc.gov/users/data-and-file-systems/transferring-data/

Asking Questions, Reporting Problems

- **Contact NERSC Consulting**
  - Toll-free 800-666-3772
  - 510-486-8611, #3
  - Email *consult@nersc.gov*.
  - https://www.nersc.gov/users/getting-help/

Thank you.

NERSC File Systems



File System Suggestions

- **DO NOT RUN BATCH JOBS IN $HOME**
  - Use $SCRATCH for running Edison/Hopper batch
  - Use $GSCRATCH for running Carver batch
- **Performance can be limited by metadata**
  - Do not store 1000s of files in single directory
- **Use "tar" to conserve inodes**
- **Use HPSS to archive important data**
  - Protection against hardware failure
  - Quota management
- **DO NOT USE /tmp!**

Local File Systems on Hopper

- **Hopper *scratch* file systems**

  `/scratch`

  `/scratch2`

  - **Each has 1.0 PB**
  - **Each has 35 GB/s aggregate bandwidth**
- **`Provided by Cray, based on Lustre`**

Hopper Scratch Use

- **Each user gets a scratch directory in /scratch1 *and* /scratch2**

  `/scratch/scratchdirs/dpturner`

  – `$SCRATCH`

  `/scratch2/scratchdirs/dpturner`

  – `$SCRATCH2`

- **Tuned for large streaming file access**
  – Running I/O intensive batch jobs
  – Data analysis/visualization

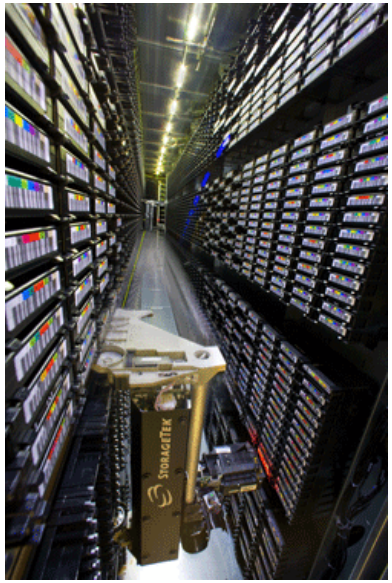
Hopper Scratch Policies

- **Quotas enforced by submit filter**
  - Combined (scratch/scratch2) quotas
  - 5 TB
  - 5,000,000 inodes
  - Quota increases may be requested
  - Monitor with **`myquota`** command
- ***Temporary* storage**
  - Daily purges of *all* files that have not been accessed in over 12 weeks
    - List of purged files in $SCRATCH/purged.<timestamp>
  - Hardware failures and human errors *will* happen

BACK UP YOUR FILES TO HPSS!

Reading from Tape

- **Loading a tape into a drive is one of the slowest system activities**
- **Positioning a tape to the beginning of data is slow compared to seeking on disk**
  - Reading a few large files from tape is relatively quick
  - Reading many small files stored on multiple tapes is slow
- **Minimize tape mounts and positioning activity for best read performance**

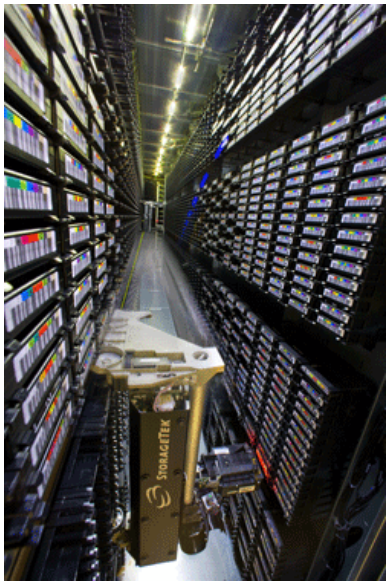


Globus Issues

- **Retry Logic**
  - GO retries failed transfers until they succeed
  - Transfers that fail for non-transient issues (e.g. permissions, quota) show up as repeated HPSS errors
    - Can lead to administrative action
- **Recursive directory syncs**
  - Can store a lot of small files—Use tar or HTAR
- **Interrupted writes to HPSS**
  - Resume not possible with current interface—interrupted transfers start over from the beginning
- **High-latency/unreliable networks**
  - HPSS very sensitive to transfer failures. Store to NGF first if using unreliable connection

HPSS is a Shared Storage Resource

- **No exclusive access to the archive**
  - No batch system
  - Inefficient use affects performance for everyone
- **The archive relies on mechanical devices**
  - Robots, tape drives, tape cartridges
  - Limited number of drives and robots to serve requests
- **Avoid administrative action**
  - Group small files together/avoid excessive writes
  - Order your retrievals
  - Excessive transfer failures (Globus with unreliable network)



NERSC Data Resources

- **Two types of files systems**
  - 'Temporary' storage, fast IO
    - Named scratch
    - Good for output from batch jobs
    - Can be either local to the machine or global to all of NERSC
  - 'Permanent' storage, slower IO
    - Named project
    - Good for longer term storage of data that is being actively analyzed
    - Visible on all NERSC systems
- **Long Term Archive**
  - Tape based, named HPSS