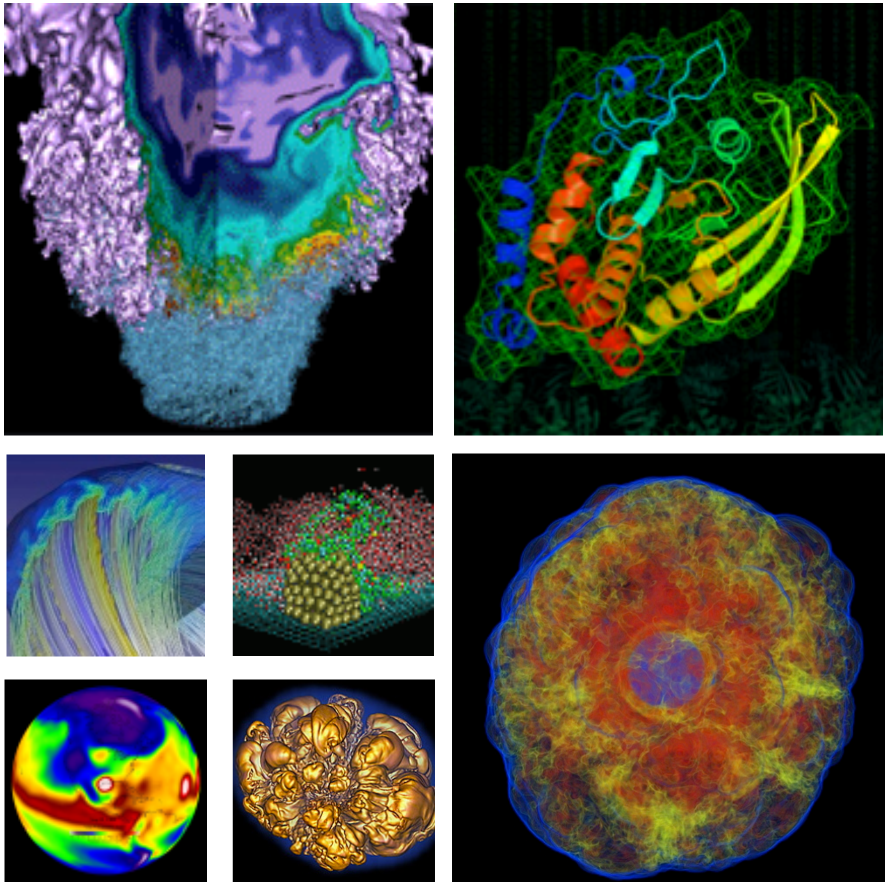
NERSC Archival Storage: Best Practices



Lisa Gerhardt

NERSC User Services

Nick Balthaser

NERSC Storage Systems

Joint Facilities User Forum on Data Intensive Computing

June 18, 2014

Agenda

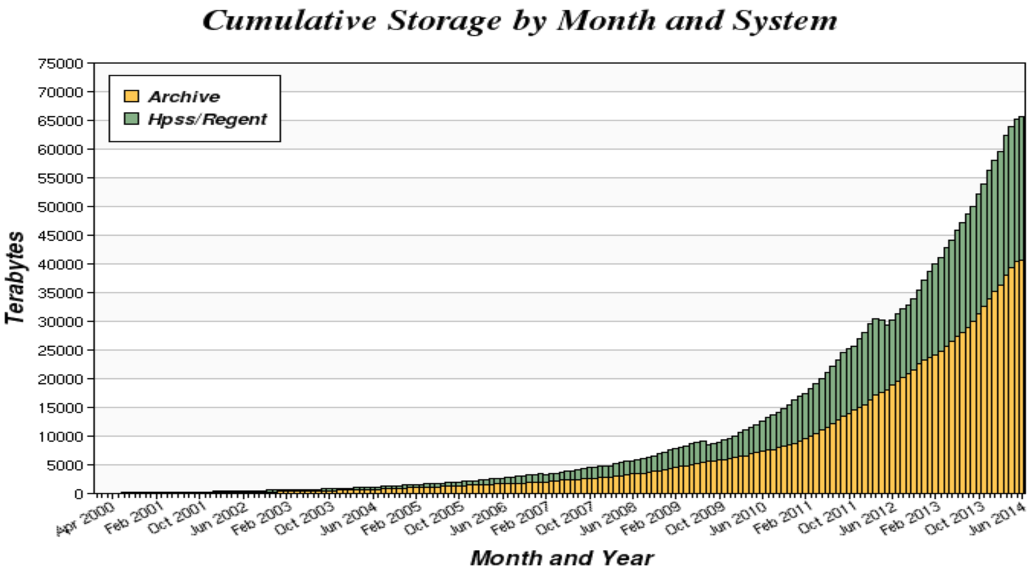
- **Introduction to Archival Storage**
  - Archive vs. backup
  - Data lifecycle management and tiered storage
  - Features of the NERSC archive
  - NERSC user case study
  - Access methods (clients)
- **Optimizing Archival Storage**
  - Tape IO characteristics
  - Storage and retrieval strategies
- **Data Redundancy and Integrity**
  - 3-2-1 Rule
  - Checksums

What is an archive?

- **Long-term data storage**
  - Data that is no longer modified or regularly accessed
  - Often the only copy of the data
  - Valuable data to be kept for the long-term
- **An archive is not a backup**
  - A backup is a copy of production data
    - If a backup is lost, production data is still online
  - Value and retention of backup data is short-term
  - Main purpose of a backup is fast recovery
- **NERSC archive has files dating back to the 1970s**
  - NERSC began using HPSS storage system software in 1998
  - Data migrated from previous archive systems including CFS and Unitree

Why should I use an archive?

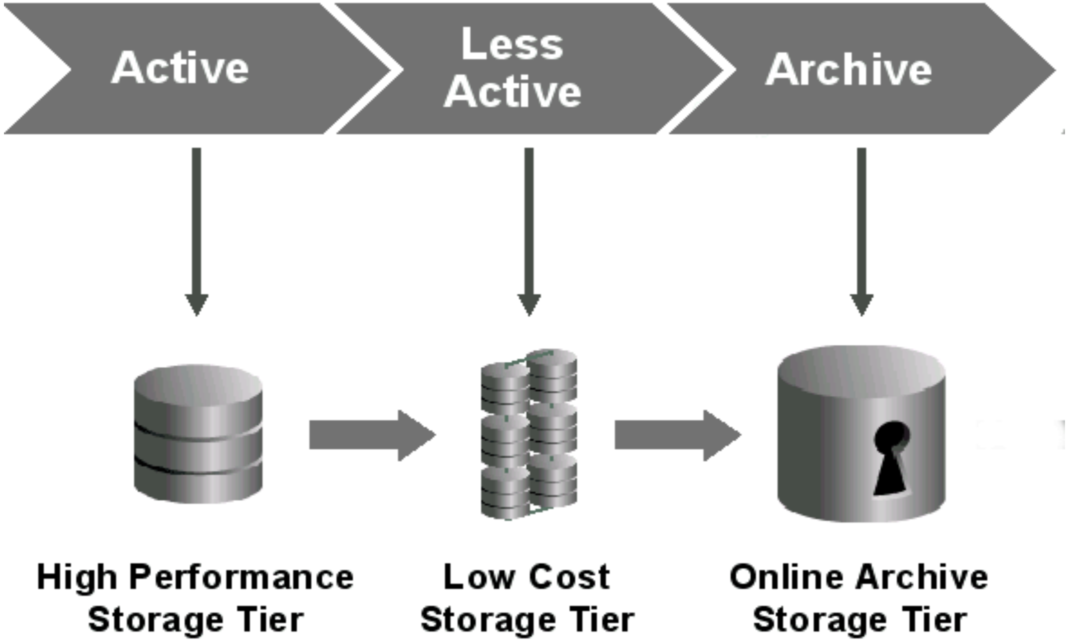
- **Data growth is exponential**



- **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
  - Important, but less frequently accessed data should be stored in an archive to free higher performing resources for processing workload

Data Lifecycle Management

- **Manage data according to access patterns and media cost**



At NERSC

Scratch File System
High capacity, fast access, high IO throughput
Temporary storage of large data files and compute output
Regular purges, not backed up

Project File System
Medium capacity, medium-term storage of shared data
Designed for file sharing

HPSS
Tape-backed storage system

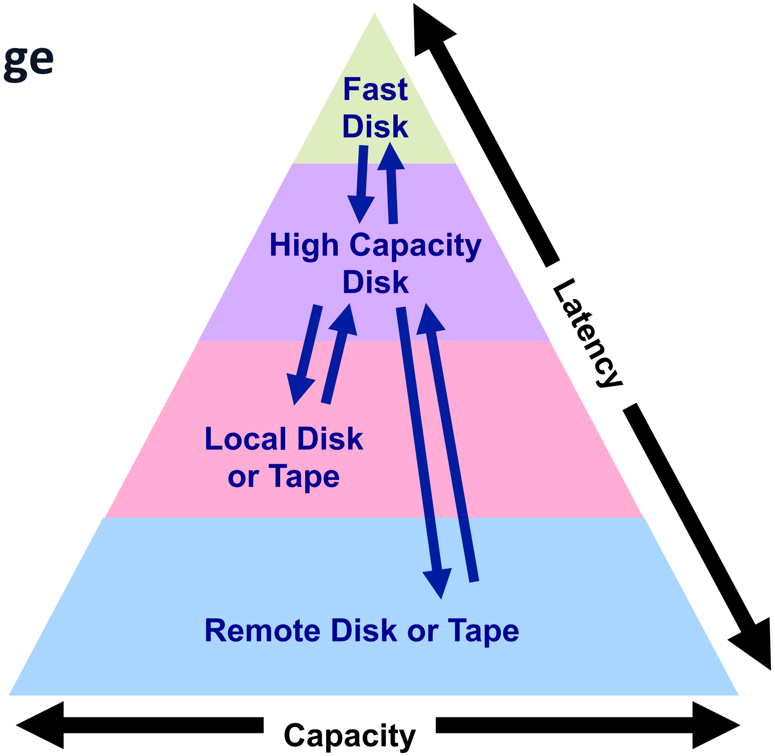
NERSC users are responsible for their own data management

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 on 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 a shared multi-user system**
  - **No batch system**. Inefficient use affects others.
  - Session limits enforced

The Archive is an HSM

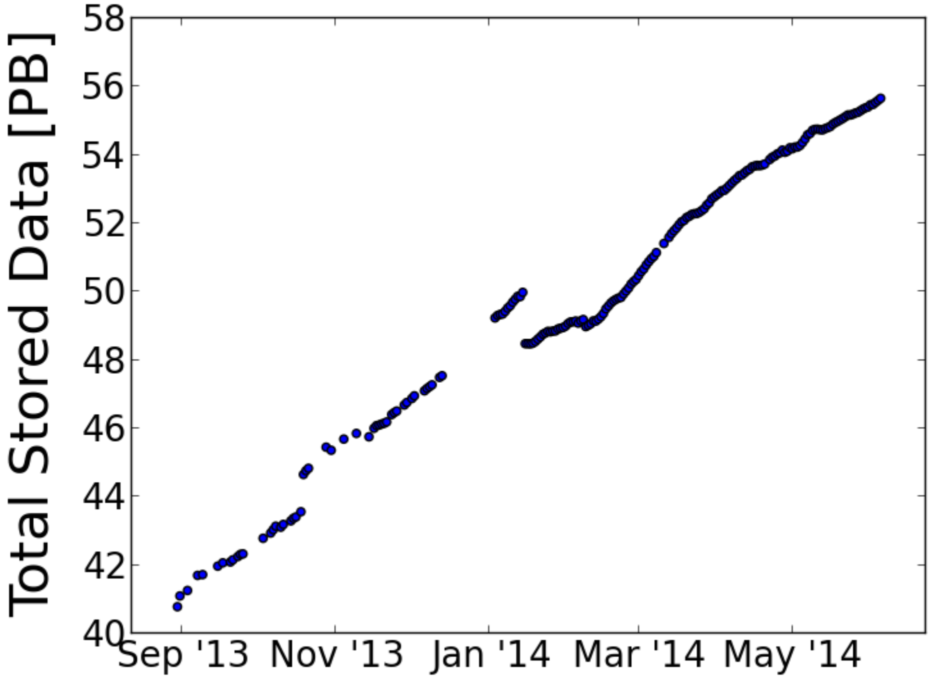
- The NERSC archive is a Hierarchical Storage Management system (HSM)
- Highest performance requirements and access characteristics at top level
- Lowest cost, greatest capacity at lower levels
- Migration between levels is automatic based on access patterns



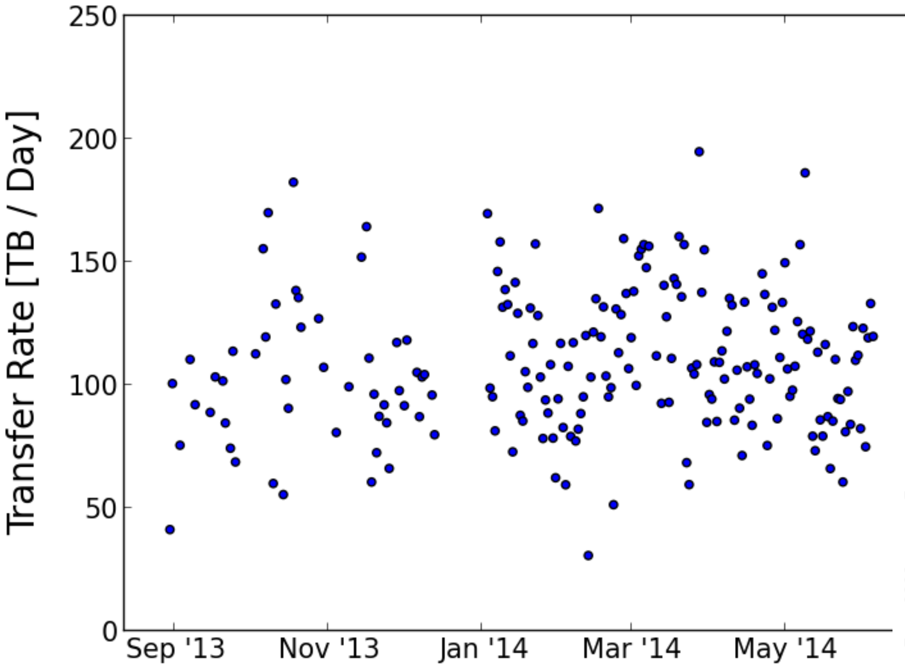
- HPSS performs differently than a file system
  - Metadata in transactional DB (IBM DB2) for integrity
  - Time to move data between hierarchies

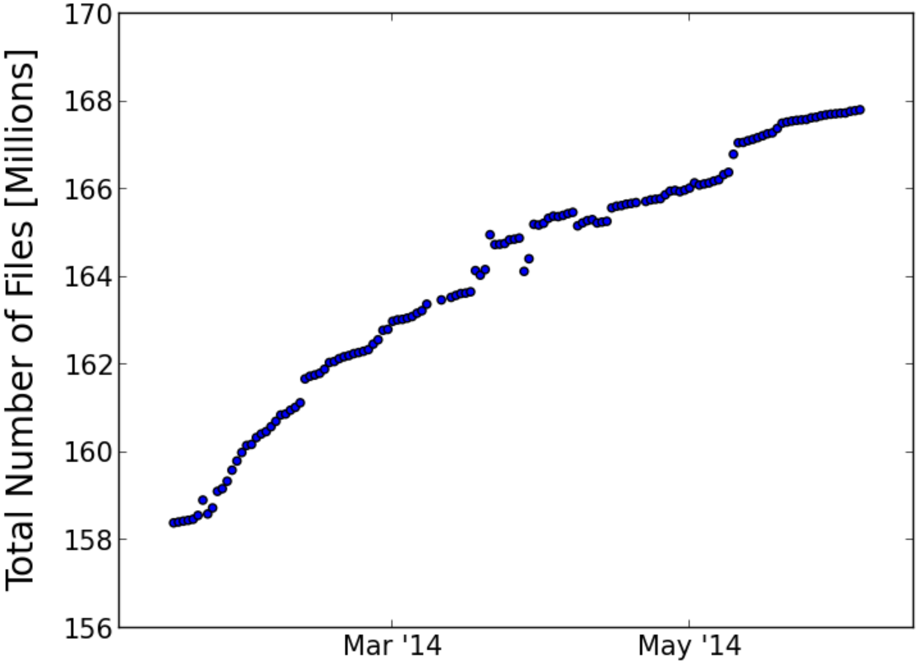
NERSC Archive: HPSS

Currently holds about 60 PB. Total capacity is 240 PB.



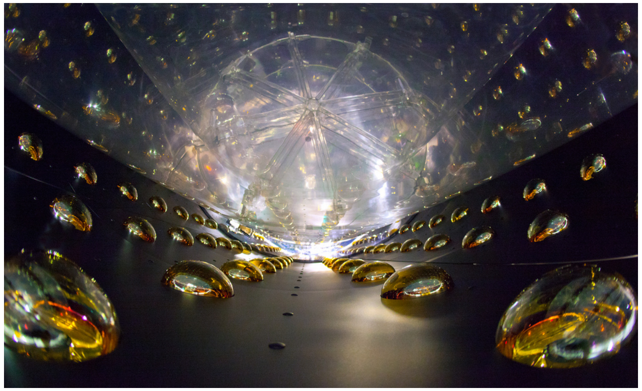
HPSS is Heavily Used



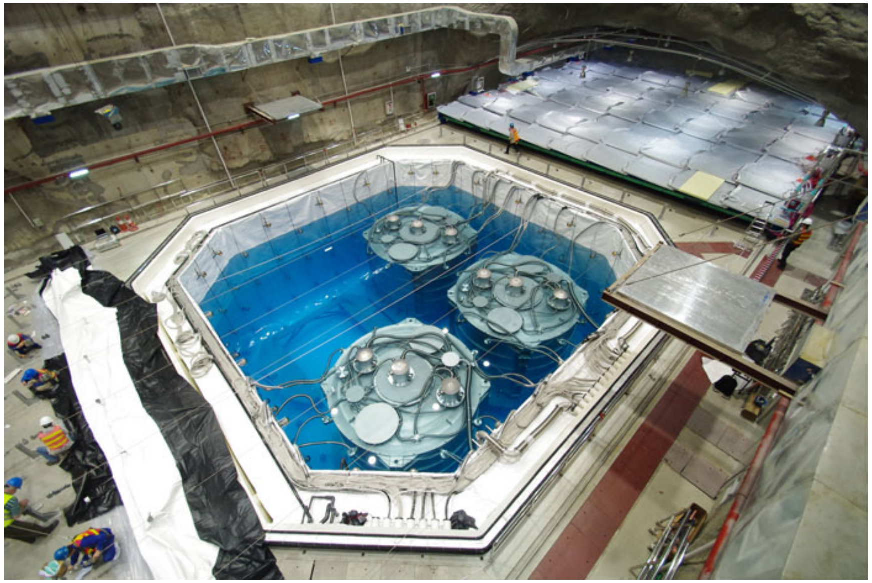


NERSC User Case Study: Daya Bay

- **Neutrino oscillation experiment in China**
- **High precision measurement of neutrino oscillation parameter**
  - Science Magazine's top ten breakthroughs of 2012
- **Data analysis and simulation done primarily on PDSF (HEP / NP cluster) at NERSC**

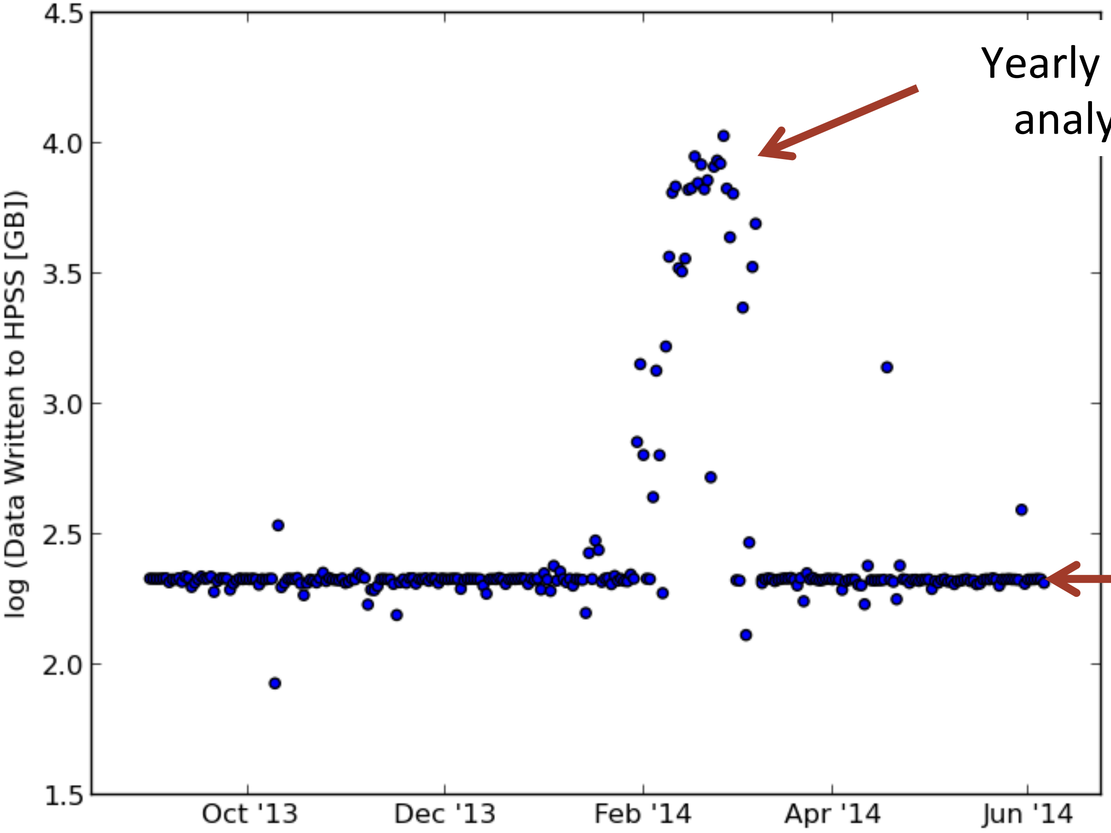


Science AAAS



NERSC User Case Study: Daya Bay

- **NERSC is their US Tier 1 facility**
  - Archive of raw data on HPSS
- **Data copied to HPSS within 15 – 20 min. of a run finishing**



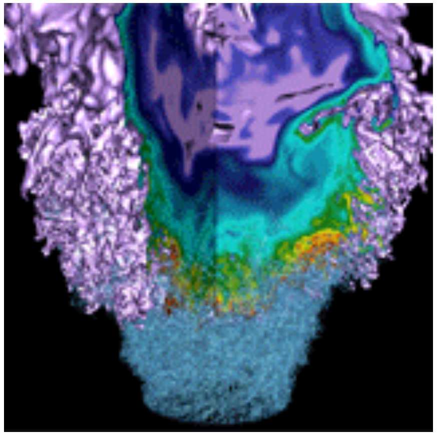
Yearly data analysis

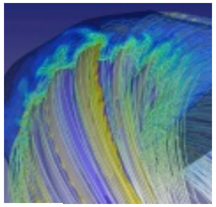
Constant upload rate of about 220 GB / day

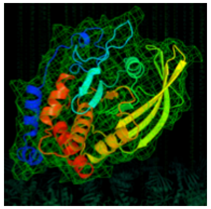
Accessing NERSC HPSS

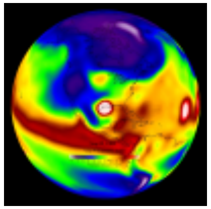
- **HSI**
  - Fast, parallel transfers, unix shell-like interface
- **HTAR**
  - Parallel, put tar file directly into HPSS, excellent for groups of small files
- **FTP/PFTP**
  - Standard and high-performance parallel FTP (NERSC platforms)
- **gridFTP**
  - Grid (GSI) authentication
  - Enables 3rd-party transfer
- **Globus**
  - Web-enabled reliable transfer

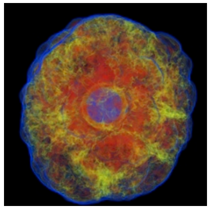
Optimizing Archival Storage

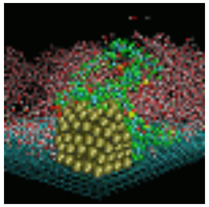












NERSC 40 YEARS at the FOREFRONT

U.S. DEPARTMENT OF ENERGY | Office of Science

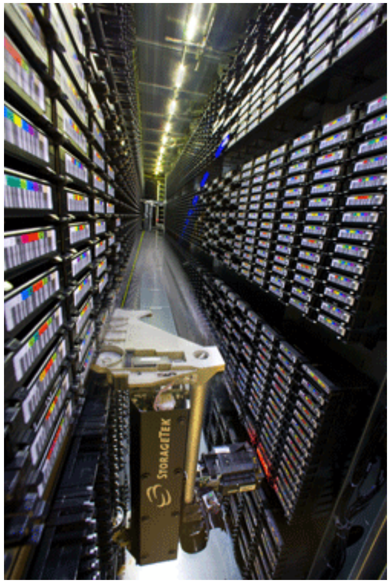
BERKELEY LAB
Lawrence Berkeley National Laboratory

Tape IO Characteristics

- **Ultimately all data in the archive is written to tape**
- **Tape is sequential (linear) media**
  - Behaves differently than disk media:
    - Data cannot be re-written in place, it is appended after the end of existing data
    - 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 pause after reading or writing each file (file sync)
- **HPSS does not respond like a normal file system**
  - Presents itself as one, but some operations can have unexpected results

Reading from Tape

- **Loading a tape into a drive and positioning to the beginning of data are the slowest system activities**
  - Average time to data on current drive technology is 45s, +15 – 30 sec cartridge load
  - Reading a few large files from tape is relatively quick – up to 400MB/sec
  - Reading many small files stored on multiple tapes is slow
- **Minimize tape mounts and positioning activity for best read performance**



Size Matters

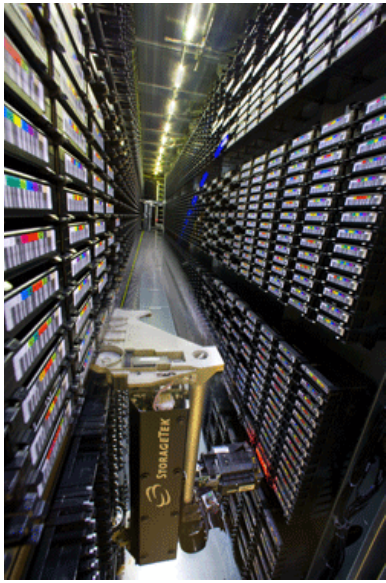
- **Sweet Spot**
  - Tape drives perform best when operating at full rate ("streaming") for long durations
  - Large file IO is best for tape drive performance
  - Small file IO causes frequent pauses and low-speed operations
  - File bundles in the **100s of GB** currently provide best performance
- **Aggregate (bundle) small files for optimal storage**
  - Use HTAR, GNU tar, and/or zip to bundle groups of small files together to optimize tape and network performance
- **There is such a thing as too big**
  - Long-running transfers can be failure-prone
  - Files spanning multiple tapes may incur tape mount delays

Tape Ordering for Non-optimal Retrieval Workloads

- **When retrieving data from the archive:**
  - **Minimize tape mounts**
    - Retrieve all files on a particular tape while it is mounted, before retrieving files on subsequent tapes
  - **Minimize tape positioning**
    - Retrieve files resident on a particular tape in order of tape position
  - **Contact NERSC Consulting for details of procedure**

Using 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**
  - Bundle small files together/avoid excessive writes
  - Order your retrievals
  - Avoid excessive transfer failures



Data Redundancy

- **Use the "3-2-1" Rule for critical data:**
  - **3** copies of the data (in different places)
  - **2** different formats
  - **1** copy off-site
- **Data loss in the NERSC archive is rare but it can happen:**
  - No offsite backup for NERSC archive data
  - Archive data is single-copy by default
  - No "Trash Can" – deleted files are gone!
- **Make multiple copies of data you care about**
  - If off-site copies are not possible at least store multiple archive copies several days apart (different tapes)
    - Contact NERSC Consulting if dual-copy is a persistent requirement (data charges apply)
- **We take great care to preserve archival data**
  - But data loss can still happen!

Data Integrity

- **Corruption**
  - Any unintended change to data
  - Rare but it can happen
    - any point during reading, writing, transfer, or processing is subject to unintended change
    - Bad network interface, disk controller, computer memory, etc. at any point in data lifetime can introduce an error
  - Corruption is data loss—make multiple copies
- **Use checksums for critical data**
  - Record checksums before storing critical data in the archive, check after retrieving
  - Many checksum methods including HSI options
  - **Checksums incur performance penalty** due to CPU load
    - check with system administrators before running checksums
- **Interrupted/failed transfers are not data corruption**
  - Check transfer return codes!

Asking Questions, Reporting Problems

- **Contact NERSC Consulting**
  - Toll-free 800-666-3772
  - 510-486-8611, #3
  - Email *consult@nersc.gov*.

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*

NERSC 40 YEARS at the FOREFRONT

Thank you.

Tape Ordering for Non-optimal Workloads

- **Minimize tape mounts**
  - Retrieve all files on a particular tape while it is mounted, before retrieving files on subsequent tapes
- **Minimize tape positioning**
  - Retrieve files resident on a particular tape in order of tape position
- **Find volume name and tape position for every file:**
  - HSI “*ls -X*” or “*ls -P*” arguments:

```
A:/home/n/nickb-> ls -P z.tar z.tar.idx
FILE   /home/n/nickb/z.tar    464566784    464566784    3095+0  EP251400    5    0    1    07/03/2013    15:00:57
07/03/2013    15:01:01
FILE   /home/n/nickb/z.tar.idx      90912         90912      608+0  EF202900    4    0    1    07/03/2013    15:01:01
07/03/2013    15:01:02
```

Position Volume

Retrieve Files in Tape and Position Order

- **Generate per-volume file lists in tape position order**
  - Using HSI, generate lists of files per tape and sort in ascending position
  - Put file path names in HSI command file using HSI “get” syntax:

```
bash$  cat ./EF2092.cmd
get z.tar.idx : /home/n/nickb/z.tar.idx
get x.tar : /home/n/nickb/x.tar
get x.tar.idx : /home/n/nickb/x.tar.idx
quit
```

- **Retrieve files in per-volume lists using HSI command file**
  - Use HSI “in cmd_file” syntax:

```
hsi –q "in ./EF2092.cmd“
Username: nickb  UID: 33065  Acct: 33065(33065) Copies: 1 Firewall: on [hsi.4.0.1.2 Thu Oct 25 16:31:52 PDT
2012][V4.0.1.2_2012_10_22.02]
get z.tar.idx : /home/n/nickb/z.tar.idx
get  'z.tar.idx' : '/home/n/nickb/z.tar.idx' (2013/07/03 15:01:02 90912 bytes, 1523.1 KBS )
get x.tar : /home/n/nickb/x.tar
get  'x.tar' : '/home/n/nickb/x.tar' (2013/07/05 13:57:06 63488 bytes, 1563.4 KBS )
```

- **Contact NERSC Consulting for full procedure**