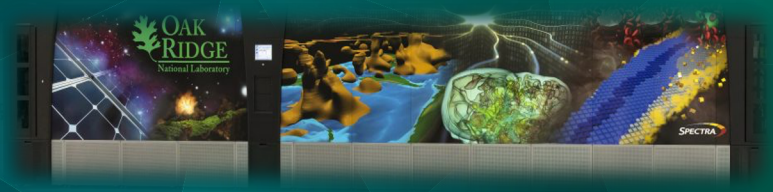
OAK RIDGE National Laboratory

HPSS Storage Archive Overview



Presented by: Gregg Gawinski

HPC Storage and Archive, National Center for Computational Science

ORNL is managed by UT-Battelle LLC
for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

What's Covered

- HPSS Terminology
- Separate Classes of Service (COS)
- Archive Overview
- File Operations
- Architecture
- Hardware Breakout
- User Interfaces
- Challenges
- Futures

HPSS Terminology

- Storage Class
  - Disk or Tape
  - Describes the storage technology and some attributes (size of a physical volume, disk or tape, type of tape, number of stripes, etc.)
- Hierarchy
  - Defines how data sits on desired storage classes
  - Disk-to-Tape (all NCCS production hierarchies are disk-to-tape)
- Class Of Service (COS)
  - Define characteristics on when data is staged to disk, min file size, max file size, etc.

Separate Classes of Service (COS)

- ORNL has four Classes of Service
  - Small less than 16MB
  - Medium 16MB to 8 gigabyte (GB)
  - Large 8GB to 1 terabyte (TB)
  - Extra-large 1TB to 256TB

    ** largest file in the archive currently is **112TB**
- We separate the Classes of Service to account for performance differences between small, medium, large, and extra-large files.
- Each COS has a different stripe width of 1-, 2-, 3-, and 6-way stripe, to try to achieve a better single stream performance per COS.
- We do not need a 1MB file to have a transfer rate of 10GB+ per second (/s); however, we do need that rate for 1TB+ files.

Separate Classes of Service (Continued)

- Most of the file count in the archive are small files less than 16MB. The average size of those files is 1.3MB and there are approximately 440 million (M). However, most of the capacity is taken up by the large and extra-large files.
- In the ORNL archive we try to keep all small files on disk and as many large files as possible, providing an increase in retrieval performance. We only purge files in the small file COS when space usage is over 90% and over 85% in the large file COS's.
- We use triple copy tape protection for our small files, since the amount of data is relatively small, it doesn't cost us many tapes to add a third layer of protection.
- Our larger files use a Redundant Array of Independent Tape (RAIT) 3+1 configuration, which provides the performance of combining four tape drives to transfer the larger files and data protection using striping and parity.

Archive Overview

- The HPSS Storage Archive consists of 2 DDN disk caches, one for small files, one for all other files (medium, large, and x-large) and a Spectra Logic tape library.
- The HPSS Archive is designed to be a cold archive with a warm cache; all files are copied to tape. The disk cache is only large enough to contain a percentage of the data based on file age. Some recalls require staging the data back to disk before it can be read.
- The user interfaces available for interaction with data in the archive are HSI, HTAR, Globus and in the future HPSSFS-FUSE.
- The archive currently contains 613,165,882 files and the used capacity is 116 PB.
- Users gain access to the HPSS archive if their project is a OLCF category 1 moderate data project.

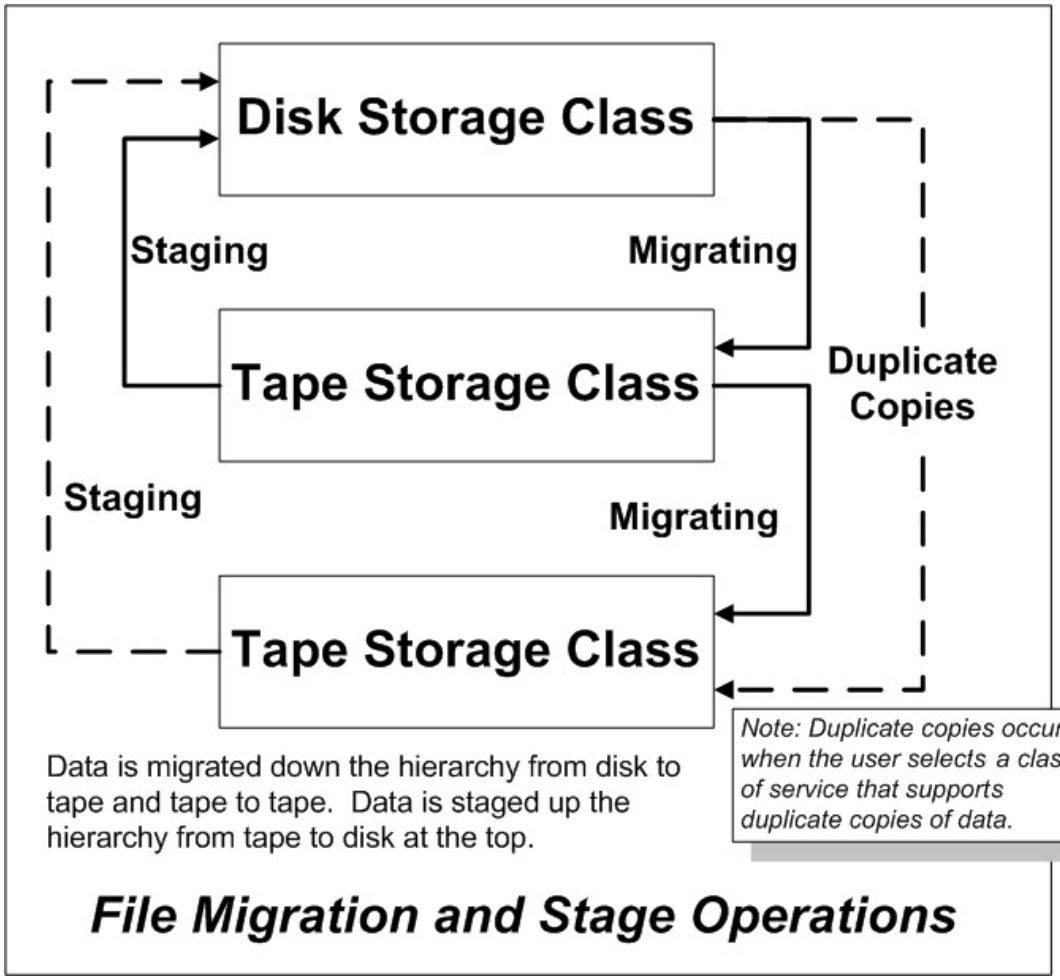
Archive Overview (Continued)

- The archive file system is setup to mirror the scratch file system, users have a symlink to their project folder in /home/<username>/<project>/ with a proj-shared, users, and world-shared directories.
- When a user puts data into HPSS, the client (HSI, HTAR, Globus) tells HPSS the size of the file, and HPSS creates the file in the correct class of service. The class of service determines which disk cache the file is then copied too. If it is <16MB it goes into the small file disk cache and if it is >16MB is goes into the large file disk cache.
- Once in the disk cache, HPSS begins migrating a copy of the file to tape in the background based on the migration policy, which the user does not see. Once the copy to tape is complete, the file is eligible for purge, based on its last accessed time.
- If the disk cache hits the configured purge policy threshold, the oldest last access time files will be purged first until the capacity drops below the threshold.

Archive Overview (Continued)

- Because the archive disk cache has a finite amount of space and it is very expensive, data needs to be purged from disk periodically leaving it on tape only. When data is recalled from tape only, users experience relatively slow performance.
- Once a user recalls data that is on tape only, that data is staged back to the disk cache for the user to access it. At this point, the file again becomes eligible for purge to tape only.
- This is the reason that users can, at times have slow retrievals and then fast retrievals depending on what medium the data currently lives on.
- We are not currently using color coding or specific file locations to identify data lives only on tape. A user needs to do a HSI ls –U if they wish to see what medium the data is on.
- Users should be aware that data being recalled from tape can take many minutes, hours or even days depending on size and order in the queue. RAIT tape sets can only transfer data at 1.2GB/s maximum, single tape sets at 360MB/s.

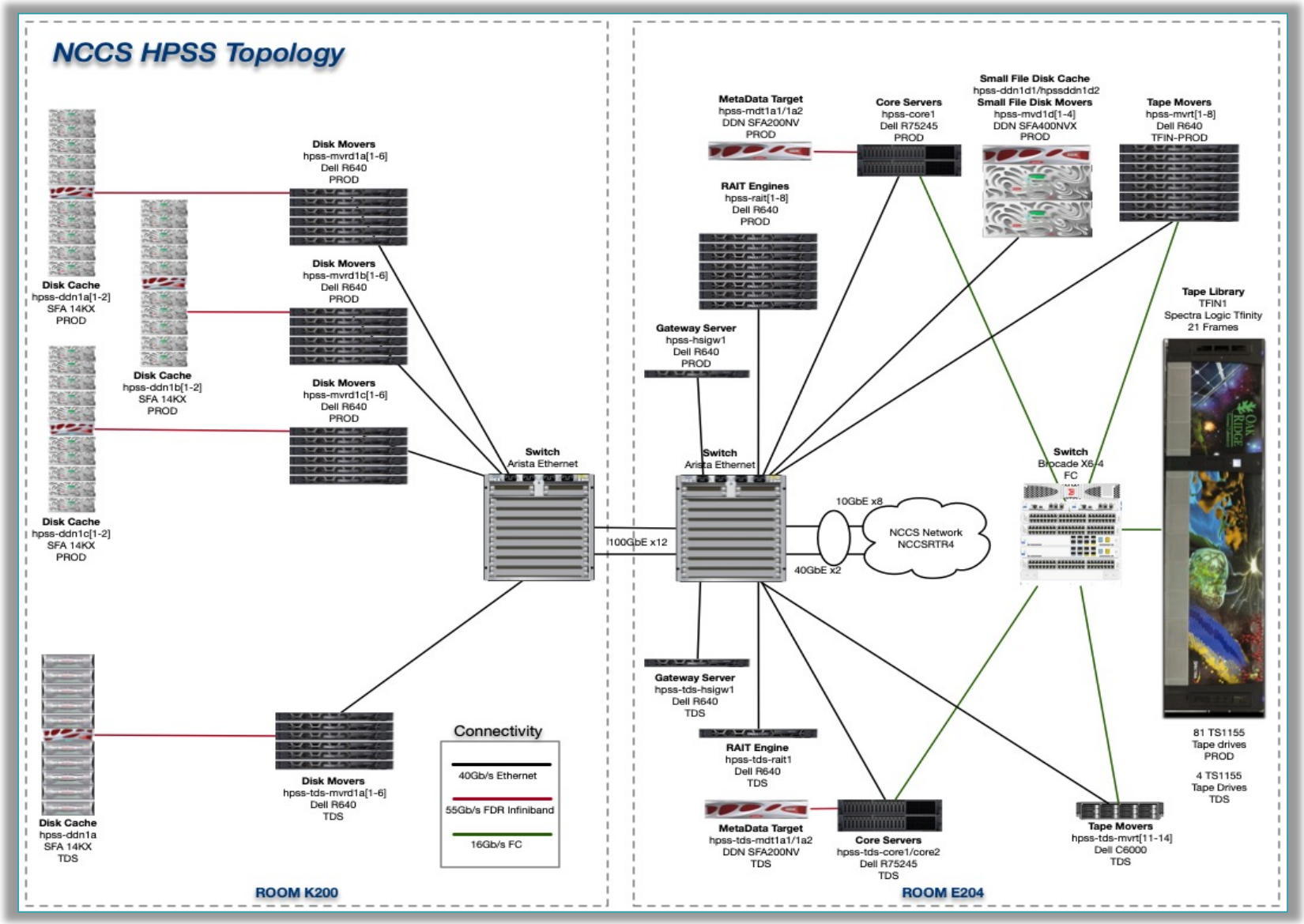
File Operations



https://www.hpss-collaboration.org/documents/HPSS_7.5.3_Installation_Guide.pdf

Figure 2.3. HPSS components

Architecture



Hardware Breakout

- Spectra Logic Tfinity ExaScale Tape Library
  - 21 Frames, 51feet long, 24 thousand pounds.
  - 81 IBM TS1155 Tape Drives (360MB/s)
  - 15,673 Tapes at 15TB per tape of capacity
  - 177.23PB total capacity in current configuration
  - 65% of capacity is in use, 61PB of space unused
  - 2 Robots on a linear rail
  - Robotic Transit Times 300ms
  - Tape Drive Mount Times ~7 seconds

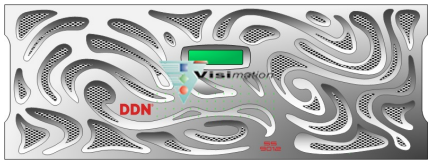
(Mounting is the process of collecting a cartridge from a storage slot, then moving it to the drive and loading the cartridge into the drive.)

Hardware Breakout (Continued)

- Small File Disk Cache

  - DDN SFA400NVX
  - Capacity ~975 TB
  - single file transfer to a single LUN speeds at 242MB/s read and 264MB/s write

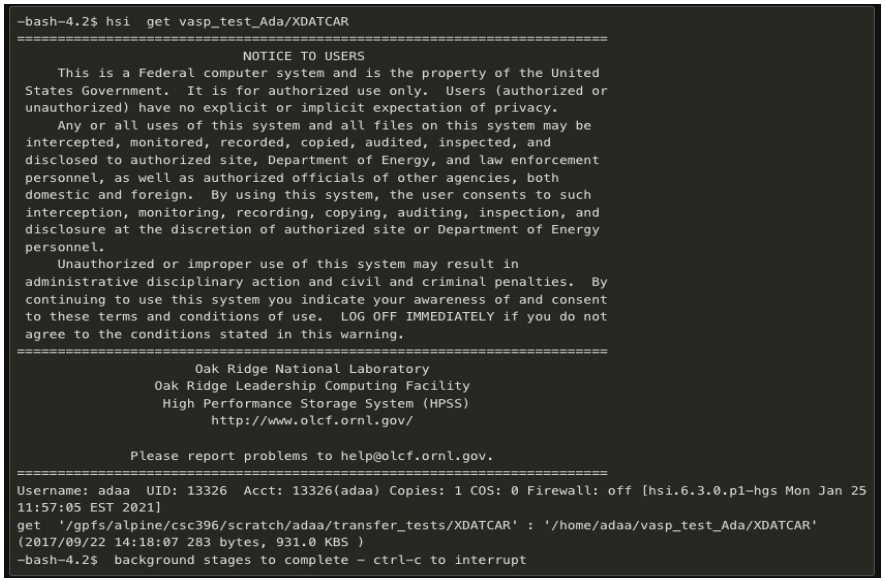
- Disk Cache



  - DDN SFA14KX
  - Capacity ~22 PB
  - single file transfer to a single LUN speeds at 3.5GB/s read, and 4GB/s write

User Interfaces

- HSI/HTAR
  - Users should be aware that recalls can take a long time, if you are using the HSI as command line and not interactive session, your prompt may not return right away. Because the press ctrl-c is on your command line, it can look as if it is frozen.



  - Bundle your files if you can, larger transfers stream better to HPSS and recall better from tape.
  - Using a DTN for data transfer is best practice.

User Interfaces (Continued)

- Optimal way to retrieve a set of files using HSI
  - https://www.hpss-collaboration.org/documents/HSI_9.3_Reference_Manual.pdf

    15.2 How to retrieve a set of files in optimal tape/position order

    The easiest way to optimize a single get command is to use the "here-document" form of the command, by creating an IN file that looks like this:

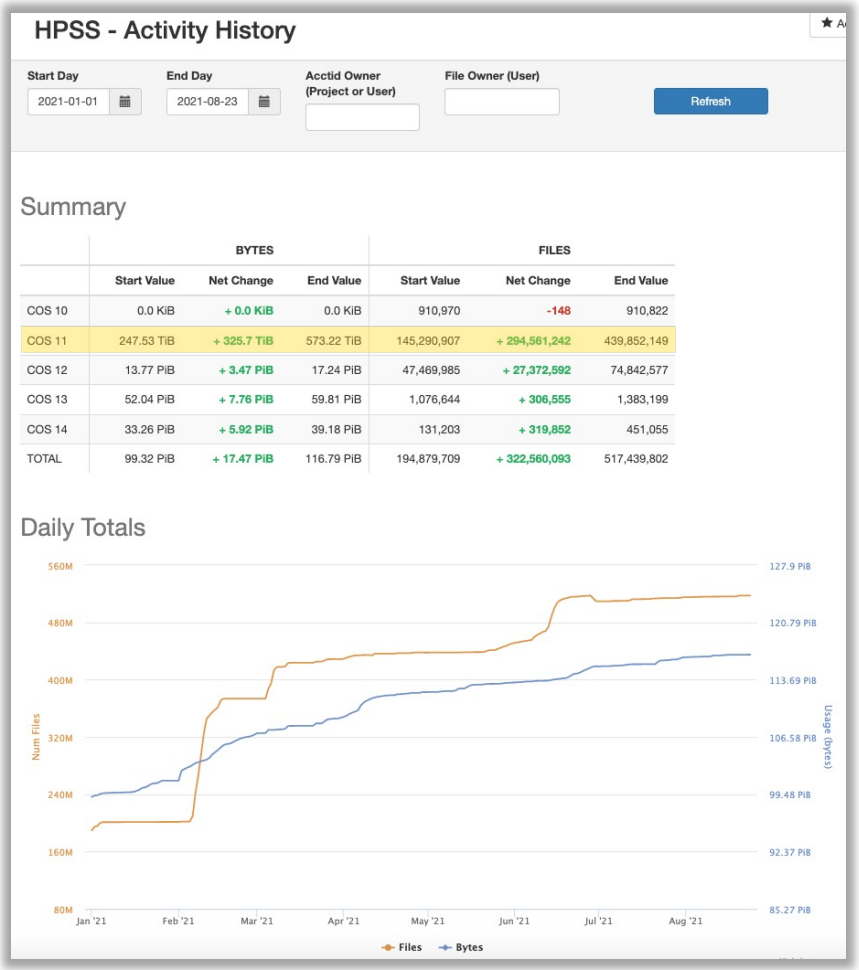
```
get << EOF
file1
file2
file3
...
EOF
```

User Interfaces (Continued)

- Globus
  - Bundle your files if you can with TAR or ZIP on a DTN node, then transfer using globus. Larger transfers stream better to HPSS and recall better from tape. Globus does not have a utility for doing this automatically.
- FUSE
  - FUSE appears as a mount on a server. Because of this, it is very simple to copy or move files over as they are. However, if the files are less than 16MB in size, they will end up on the small file disk cache where performance is slowest. For better performance, combine small files with TAR prior to moving them (just like you would in a normal HSI command).

Challenges

- COS11(Small Files) fastest growing COS
  - ~150 million files in past year
- Increased Globus usage leading to small file transfers
- For some users, tar files are not practical for retrieval
- Increasing performance has led to increased usage.



Challenges(Continued)

- Globus is a convenient interface; however, it does not provide the ability to combine many small files together for optimal HPSS performance. For example, we've had cases where transfers of large numbers of small files have filled up the cache in a few days, impacting all users, since the cache is shared.
- Testing HPSS 9.2
  - With this upgrade from 7.5.3, we plan to provide an HPSS Filesystem in Userspace (FUSE) interface and use auto change COS.
  - HPSSFS-FUSE adds convenience, but also has potential to cause issues from large ingests of small files as previously discussed.

Futures

- Open nearline system – hardware is onsite, deployment is underway. Specifications are; 30 PB usable tape storage, 16 TS1160 tape drives at a transfer speed of 400MB/s per drive, and 10PB usable disk storage at a transfer speed of 70GB/s.
- Moderate nearline system - currently in RFP process.
- Only data that requires a tape copy will be placed on tape. Unlike today where all data that is placed into the archive is also copied to tape.
- Use of visual attributes to help identify data that is on tape, such as a folder structure or color coding.
- Increased performance.

Conclusion

Questions/Thoughts?

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.