FY16 Archive Disk Cache Expansion

Livermore Computing User Meeting

Jerry Shoopman

Livermore Computing Division
Data Storage Group Leader

March 1, 2016

LLNL-PRES-685139

Lawrence Livermore National Laboratory

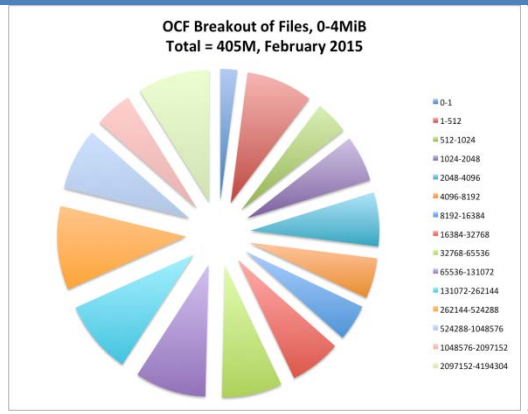
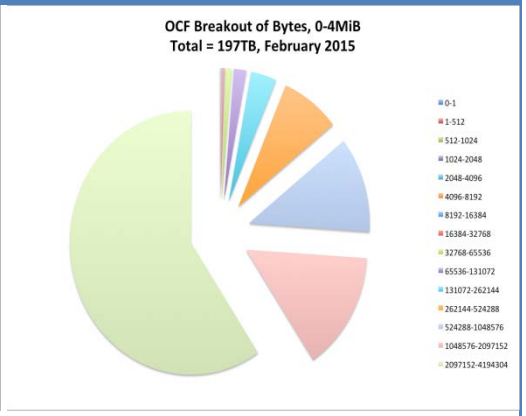
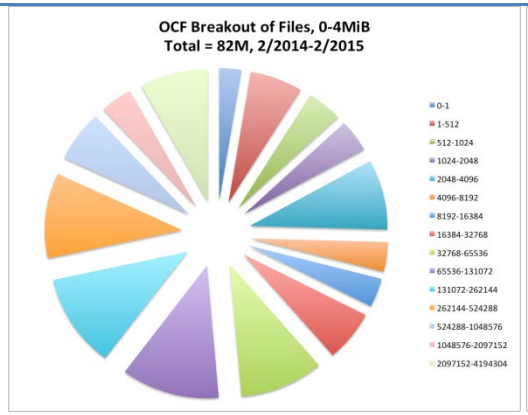
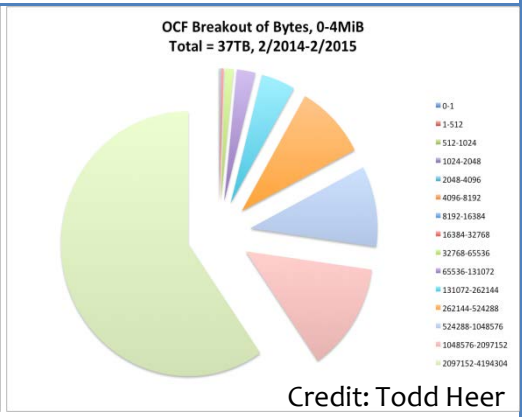
Modernization and Expansion of LLNL Archive Disk Cache – L2 Milestone 5589

- Livermore Computing completed ASC L2 milestone 5589: Modernization and Expansion of LLNL Archive Disk Cache
- **L2 Milestone Description:** *Configuration of archival disk cache systems will be modernized to reduce fragmentation, and new, higher capacity disk subsystems will be deployed. This will enhance archival disk cache capability for ASC archive users, enabling files written to the archives to remain resident on disk for many (6–12) months, regardless of file size.*
- The milestone was completed early, in three major phases:
  - August 26, 2015 subsystems w/6PB were deployed to OCF HPSS production.
  - September 23, 2015 subsystems w/9 PB were deployed to SCF HPSS production.
  - January 31, 2016, the milestone was fully satisfied when legacy archive disk cache subsystems were fully retired from production use in both LLNL's unclassified and classified HPSS environments, and only the newly deployed systems were in use.

Cross-functional LC team of Computing Operations and Networking Group, Data Storage Group, Platform Deployment Architect, System Administration Group.

Modernization and Expansion of LLNL Archive Disk Cache - Analysis

- The massive scale of LC platforms and file system expansions required the acquisition, integration, tuning, and deployment of correspondingly capable archive disk cache systems in both the OCF/SCF.
- Integrating systems within existing LC environments was critical to ASC. The target disk residency for files written to new archive disk caches was six months or more.

	OCF File Count	OCF Bytes
All files in OCF		
OCF files over one year		

Credit: Todd Heer

A thorough DSG analysis the smallest archive files was critical, as they are the most problematic. Charts above illustrate distribution of OCF HPSS files in the 0-4MB range.

Modernization and Expansion of LLNL Archive Disk Cache – Buy H/W

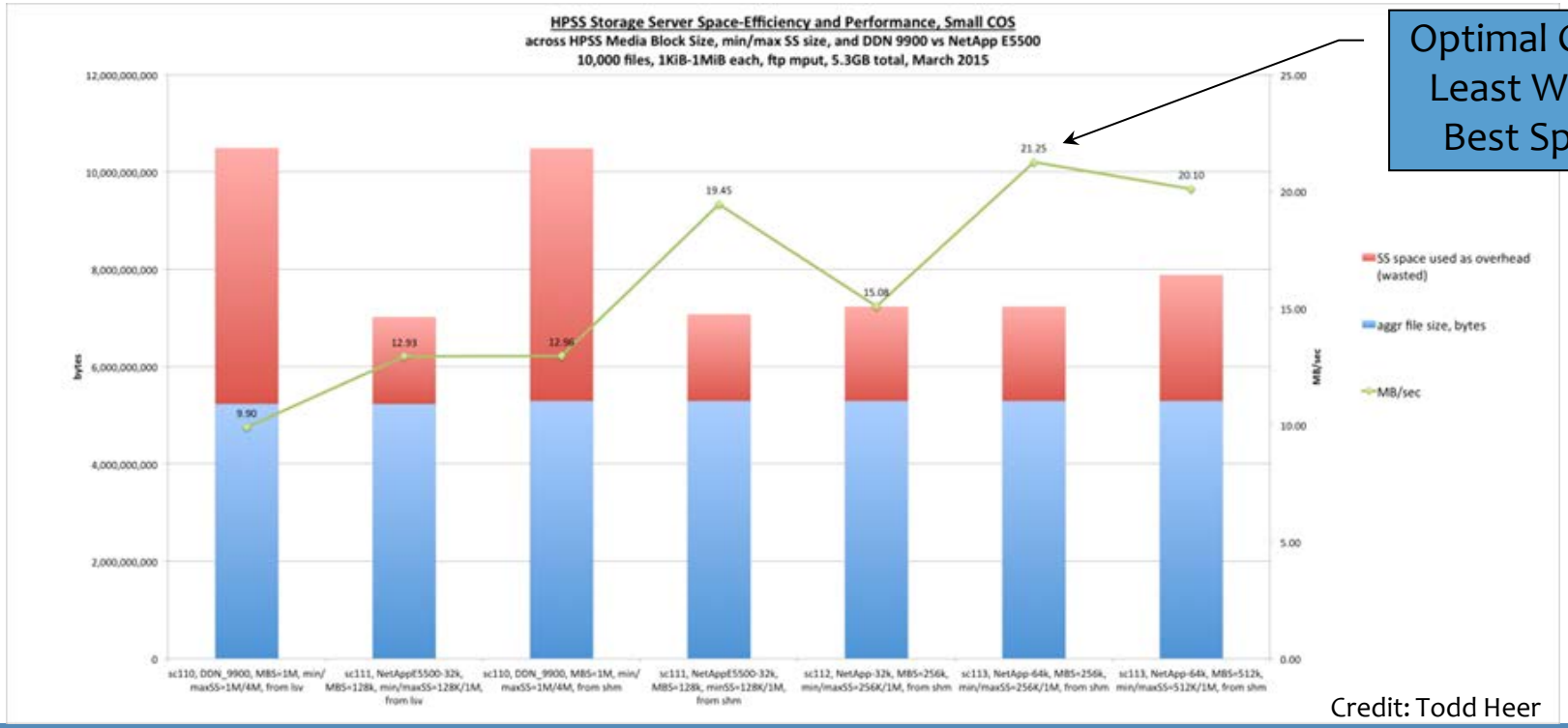
- **Seven separate procurements** were completed for new disk hardware, subsequent disk expansions, HPSS Mover nodes and node expansions, as well as 40GbE Network hardware.
- Archive disk cache systems named **Yuban** in OCF and **Osprey** in SCF were deployed.

- Yuban and Osprey archive disk cache systems consist of **80 NetApp E5500 60-disk enclosures** with **40 NetApp E5500 controller pairs & 4,800 individual 4TB hard drives**. These are hosted by **68 Haswell server nodes** w/quad-port **Qlogic FC8 Host Bus Adapters**, and **Intel 40GbE Network Interface Cards**.

The 4 RSSU OCF Yuban system is pictured. Cisco 40GbE networks collocated with Yuban and Osprey clusters, provide high-bandwidth connectivity to LC platforms.

Modernization and Expansion of LLNL Archive Disk Cache - Optimize

- To identify the optimal Small file configuration, numerous tests were performed. DSG created 10,000 files from 1KB to 1MB in size, and compared results of various low-level disk configurations.

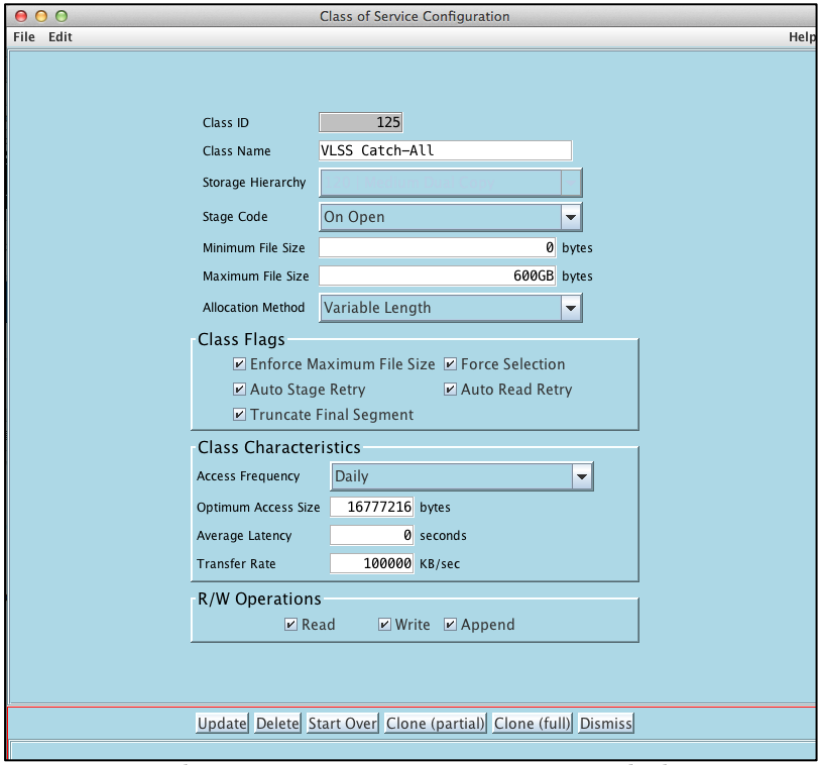


Credit: Todd Heer

DSG's scientific method: This dual-axis chart illustrates both efficiency in terms of wasted space for the given size of user bytes stored (left axis) and avg. performance for writes of 10,000 1KB-1MB files (right axis). Each bar is avg. of 4-6 test runs.

Non-LC file transfer size limits and Variable Length Storage Segments

- Traditionally, users transferring files/data to HPSS archives from outside LC networks (e.g. Enterprise Network, WCI desktops) are subject to HPSS maximum segment count of 10,000 at the minimum segment size (now 1MB). Newly created files written from such hosts could not be larger than 10GB. Previously, this was 80GB.
- LLNL HPSS system's configured maximum file size is 100TB when the client knows how to "intelligently" interface to HPSS, and alloc size.
- After testing, DSG determined that the Variable Length Storage Segment (VLSS) feature could be deployed and used for the Default COS.

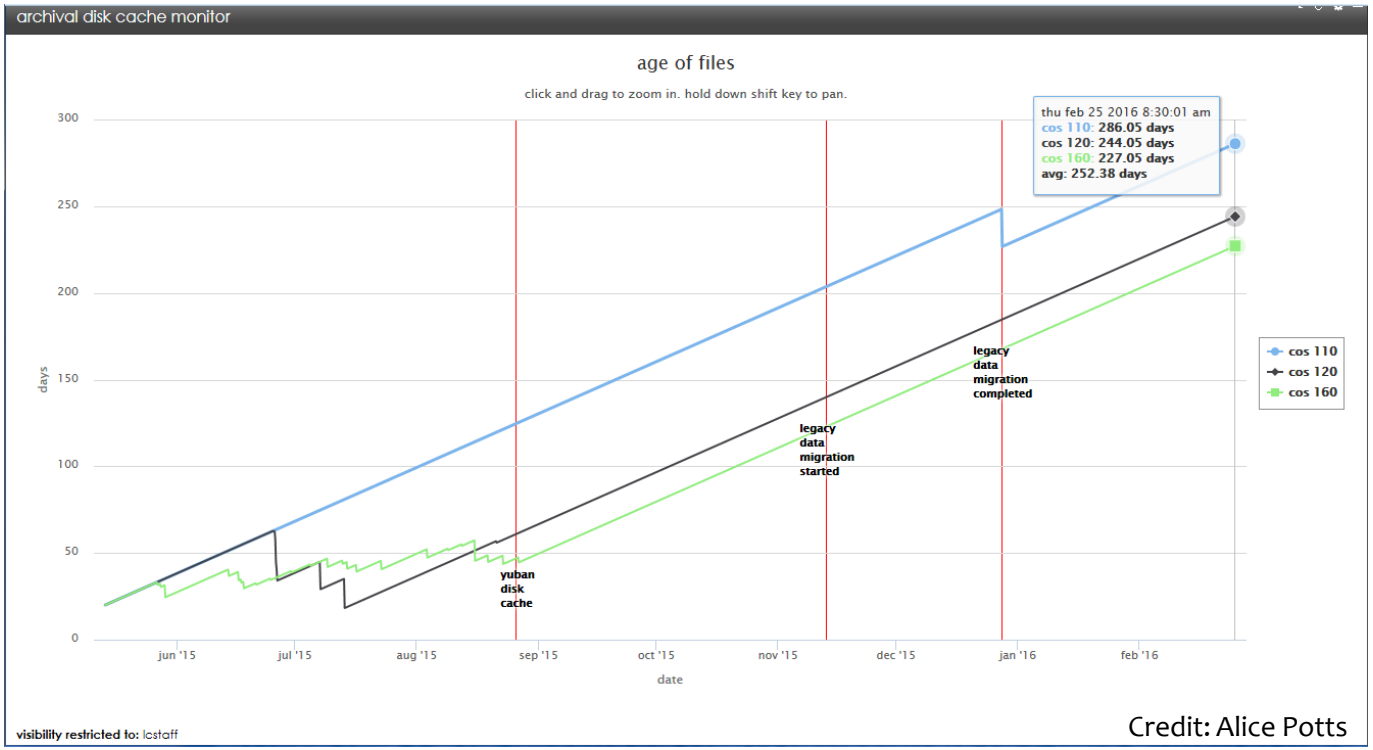


- VLSS allows HPSS to recognize files consuming segments at the min segment size, and then increase subsequent segment sizes (by powers of two) up to max seg size (64 MB for this COS/SC). VLSS allows users outside LC to write 600 GB files without special commands, avoiding the 10GB limit.

LLNL HPSS Developers enhanced the system to automatically detect large, non-LC file creates at the time of migration, modifying them to the optimal COS based on file size.

Modernization and Expansion of LLNL Archive Disk Cache - Deploy

- Hardware, firmware, and software have been tuned/optimized
- HPSS configurations have been adapted.
- Early failures have been worked to resolution.
- Legacy archive disk cache files have been migrated to new disk and/or purged from legacy systems.



Credit: Alice Potts

- HPSS Storage Class **Purge thresholds have been increased to 90%.** New storage classes vary from 7%-19% full for "small" files (0-1MB), to 73%-82% for "medium" files (1-256MB), and 50%-60% for "large" files (256MB+). Disk life residency is expected to continue to increase before reaching steady state.

Archival Disk Cache Monitor charts on MyLC demonstrate OCF data written in the last 252+ days remains resident on disk, regardless of file size. SCF average is at 155+ days.

Lawrence Livermore
National Laboratory