LA-UR-11-11388
Approved for public release; distribution is unlimited.

Title:	U.S. Department of Energy Best Practices Workshop on File Systems & Archives Position Paper
Author(s):	Torres, Aaron Scott, Cody
Intended for:	U.S. Department of Energy Best Practices Workshop on File Systems & Archives, 2011-09-26/2011-09-27 (San Francisco, California, United States)

U.S. Department of Energy Best Practices Workshop on
File Systems & Archives
San Francisco, CA
September 26-27, 2011
Position Paper

Aaron Torres
Los Alamos National Laboratory, HPC-3
agtorre@lanl.gov

Cody Scott
Los Alamos National Laboratory, HPC-3
cscott@lanl.gov

ABSTRACT / SUMMARY

As HPC archival storage needs continue to grow, we have started to look at strategies to incorporate cheaper, denser, and faster disk as a larger part of the archival storage hierarchy. The archive in Los Alamos National Laboratory's Turquoise open collaboration network has always used a generous amount of both fast and slow disk in addition to tape. Lessons learned during Road Runner Open Science pointed to the need for large amounts of cheaper, slower disk for storage of small to medium sized files and faster disk in order to store and quickly retrieve file metadata. New advances in the last year may signal another transition that altogether eliminates the need for migrating small and medium files to tape. Improvements in disk speed, particularly solid state devices (SSDs), also allow us to operate on billions of files in a reasonable span of time even as archives continue to grow.

INTRODUCTION

The Open Science simulations run on the Road Runner supercomputer at Los Alamos National Laboratory (LANL) in 2009 provided the opportunity to test an archive based on commercial off the shelf (COTS) components. For this archive, we chose the General Parallel File System (GPFS) and Tivoli Storage Manager (TSM) due to robust metadata features, fast data movement, flexible storage pool hierarchy and migration, and support for a multitude of disk and tape options [1].

This archive joins a long history of archival storage at LANL, including the Central File System (CFS) and High Performance Storage System (HPSS). Thanks to administrative diligence, we have or can recreate records about usage patterns of these archives. One similarity we keep seeing in large archives, with the COTS archive being no exception, is that we primarily store numerous small to medium sized files rather than storing large to huge files. As of May, 2011, HPSS at LANL houses nearly 163 million files with total size of 19.6 PB with an average file size of 131.5MB [2]. NERSC publishes similar statistics with an archive housing over 118 million files and 12 PB for an average size of 109MB [3].

ROAD RUNNER LESSONS LEARNED

When designing for archival storage, one often considers the extreme case for file size. In HPC this generally means designing for enormous files on the order of terabytes for current supercomputer sizes. In practice, however, we see a tremendous amount of small to medium files, especially with users performing n-to-n writes or using the Parallel Log-structured File System (PLFS) to effectively convert n-to-1 writes to n-to-n [4]. In the case of Roadrunner, 20 million 8-16 MB files were archived in one weekend [5].

For the COTS archive, this proved to be the largest pain point since the Hierarchical Storage

Manager (HSM) feature of TSM does not currently support aggregating smaller files together when moving them to tape, resulting in poor performance. Users could aggregate their own files using the "tar" command, but they cannot be relied on to do this for all cases. Another option would be to put file aggregation into an archive copy tool such as how the LANL-developed Parallel Storage Interface (PSI) does with the Gleicher developed HTAR [6]. However, doing so breaks POSIX compliance because no other standard file system tool can read or write files aggregated in this way. One of the design goals of the COTS archive was to leverage as much standard software as possible.

For the COTS archive, moving small files to tape without a transparent file aggregation technique did not make sense. So, small files are kept on RAID 6 disk arrays and backed up to tape. RAID provides recovery from minor amounts of single disk failure, and the tape backup provides disaster recovery. Moreover, TSM's backup function does support aggregating small files before sending them to tape.

The COTS archive has 122 TB of fast fiber channel disk to act as a landing area for new files and 273 TB of SATA disk for files under 8 MB to be moved to. Finally, it has 3 PB of tape for files over 8 MB and for the backups of the SATA disk pools. Currently, the archive houses over 107 million files with a total size of 2.1 PB and an average size of 21.22 MB according to our latest statistics as of August, 2011. As shown in Figure 1, 97 million files are less than or equal to 8 MB. This indicates that the general case for our archive is large amounts of smaller files.

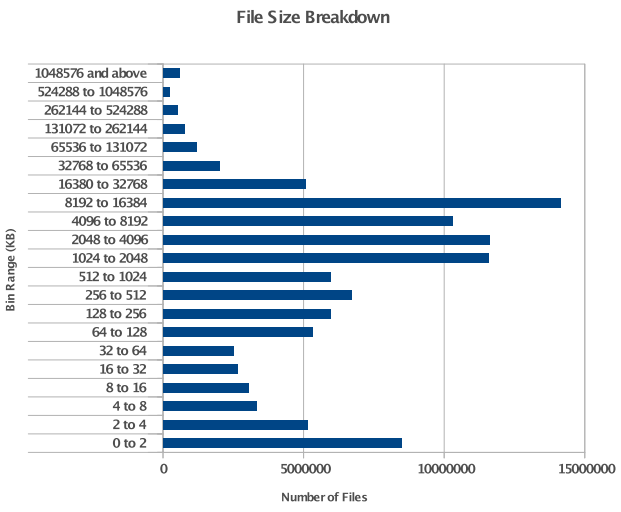

Figure 1. File Size Breakdown of COTS Archive.

RECENT ADVANCEMENTS

The recent explosion of "cloud" backup providers like Mozy, Backblaze, and others lead to questions about how we store large amounts data and if we are doing it in the most cost effective way. For a cloud-based backup service, density and uptime are the two primary driving forces because users continue to back up ever larger amounts of data as they put more of their life on the computer in terms of photos, videos, etc. and data may be backed up or restored at any time. These are also motivating factors for HPC archives. On September 1st, 2009, Backblaze posted an entry to their company blog describing their Backblaze Pod capable of storing 67 TB of data in a 4U enclosure using 47 one terabyte drives for $7,867, or 11.4¢ per gigabyte [7]. On July 20th, 2011, they posted an updated entry now indicating that they can store 135TB in 4U using 47 three terabyte drives for $7,384 or 5.3¢ per gigabyte [8]. Also, Backblaze notes they have deployed 16 PB of disk in the last 3 years [8]. In terms of raw storage, that is within striking distance of the size of LANL's largest HPSS archive at nearly 20 PB.

On the other end of the spectrum, eBay recently replaced 100 TB of SAS disk with SSD [9]. They did this to speed up virtualization and reduce the size of their disk farm. They had a 50% reduction in standard storage rack space and a 78% drop in power consumption by moving to SSD. Although it is impossible for an HPC archive to take this approach, it is possible to replace portions of the total system for tremendous benefits.

An example of using SSD in a storage hierarchy is IBM's recent efforts at speeding up GPFS using SSD [10]. By storing GPFS metadata on SSD, IBM saw a 37 times speed improvement for metadata operations and was able to scan 10 billion files in 43 minutes. For comparison, it

takes roughly 20 minutes to scan the 120 million files on the LANL COTS archive using eight 15,000 RPM fiber channel disks in four RAID 1 stripes.

CHEAP, DENSE DISK

The growth of density in hard disks shows little sign of slowing down. In the 2 years between Backblaze posting blog entries about their Pod system, the cost of the Pod actually went down even though the raw capacity of the 4U box doubled. Hard disks in the 4 TB range are on the horizon for desktops and servers in the next year [11], and even laptops are moving to 1 TB disks [12]. The Hitachi 3 TB drives used by Backblaze can be purchased for $120-130 from a variety of retailers [13]. For comparison, an LTO5 tape that holds 1.5 TB of uncompressed data costs approximately $60 [14]. For the same capacity, the disk costs as much as the tape.

One interesting move by companies like Backblaze is that they use consumer level hard drives instead of "enterprise ready" drives. Such drives are substantially cheaper; with the enterprise version of the Hitachi 3TB drives costing over $320-350 per drive as of August 23, 2011 [15]. Backblaze also takes advantage of the manufacturer's 3 year warranty to get a replacement disk if one fails rather than an expensive maintenance contract. HPC archives might be able to leverage the same kind of disk drive by taking into account the disk failure protection afforded by RAID 6 and by having a tape backup of whatever is stored on such disks.

Unlike Backblaze and their Pod, we do not want to be in the custom hardware business. So, we looked for existing commercial hardware that could get the same density of disk. We found the SuperMicro SuperChassis 847E26-RJBOD1 [16]. It is a storage chassis that can support 45 disk drives in 4U. It does not have a built-in motherboard like the Backblaze Pod to manage the disk, but the COTS archive already has machines in its GPFS cluster that can easily take a RAID card with an external SAS connector to plug into this storage expansion chassis. Filling the chassis with 3 TB consumer level disk drives and including the RAID card costs approximately $12,000, or 8¢ per gigabyte.

The idea behind these enormous disk pools is not to completely replace tape, but to adjust the size of file that gets moved to tape. It is entirely feasible today to change the threshold used in the COTS archive from 8 MB to 1 GB with the current price of disk. With this change, we can move all files 1 GB and larger to tape. This file size is also much closer to the size of file necessary to get a tape drive up to peak streaming speed based on internal testing done at LANL. In addition, backing up any file that will stay resident on a disk greatly reduces the fear of failed tapes when storing enormous quantities of small files.

One argument against disk in archive is that archives are usually "write once and read never," so it does not make sense to "waste" power and cooling on spinning disk for data that may never get read. For the COTS archive, the ability of GPFS to move data to different types of storage (ie, fast disk, slow disk, and tape using TSM) based on arbitrary criteria like dates could be leveraged to move rarely or never read data to tape. The tape performance hit is acceptable because the data is essentially "cold". Similarly, a large disk pool can be used to stage data for reading if a user knows he or she will be pulling some set of data from the archive.

FAST DISK CACHE

As HPC archives ingest ever more data because of exascale supercomputers, the metadata will probably become more and more important. At some point in the not to distant future, users will want to search the archive on metadata instead of being forced to create complex directory hierarchies to find the files they are interested in. An example query could be "find all the checkpoint files that were copied to the archive within the last 3 days." Thus, it is also important to quickly search an archive's metadata, whether it is in a file system like GPFS or a database like HPSS using DB2. Here is where faster disk systems like SSD can be used to great effect in HPC archives. As mentioned previously, IBM's

testing of storing GPFS metadata on SSD and being able to scan billions of files in less than an hour shows how such fast disk can be very useful.

Another pilot program at LANL is testing metadata performance on SSD using the GPFS COTS archive as a basis for the number of files and types of files stored. Having the ability to quickly scan the metadata of the entire archive provides many benefits, particularly to future research projects and in data management including ongoing work to index and quickly search archive metadata.

In addition, as SSD storage becomes cheaper and denser, it may eventually be possible to replace our fast disk cache, currently consisting of fiber channel disk, with a large pool of SSD similar to how eBay replaced their SAS disk environment. With our current data requirements this is still cost ineffective, but it is worth examining and testing now for the future.

CONCLUSIONS

There are many advantages to having a large, easy to manage pool of disk. When raw speed is not a requirement of this disk, there are solutions available to procure, maintain, and deploy a tremendous amount of disk cost effectively that compares very favorably to the cost of tape. Taking advantage of faster disk like SSD for metadata and disk cache will also benefit future HPC archives. The LANL COTS archive is in a unique position to test and potentially deploy some of these newer solutions in-place with limited negative effect to users.

REFERENCES

1. Hsing-bung Chen, Grider Gary, Scott Cody, Turley Milton, Torres Aaron, Sanchez Kathy, Bremer John. *Integration Experiences and Performance Studies of A COTS Parallel Archive System.* IEEE Cluster Conference, 2010
2. *Accounting Data.* http://hpss-info.lanl.gov/AcctProcess.php
3. *Storage Trends and Summaries.* http://www.nersc.gov/users/data-and-networking/hpss/storage-statistics/storage-trends/
4. Bent John, Gibson Garth, Grider Gary, McClelland Ben, Nowocznski Paul, Nunez James, Polte Milo, Wignate Meghan. *PLFS: A Checkpoint Filesystem for Parallel Applications*. Super Computing, 2009
5. Scott, Cody. *COTS Archive Lessons Learned & Fast Data Pipe Projects*. JOWOG-34
6. Gleicher, Michael. *HTAR – Introduction.* http://www.mgleicher.us/GEL/htar/
7. Nunfire, Time. *Petabytes on a Budget: How to build cheap cloud storage*. http://blog.backblaze.com/2009/09/01/petabytes-on-a-budget-how-to-build-cheap-cloud-storage/
8. Nufire, Tim. *Petabyes on a Budget v2.0: Revealing More Secrets.* http://blog.backblaze.com/2011/07/20/petabytes-on-a-budget-v2-0revealing-more-secrets/.
9. Mearian, Lucas. *Ebay attacks server virtualization with 100TB of SSD storage.* http://www.computerworld.com/s/article/9218811/EBay_attacks_server_virtualization_with_100TB_of_SSD_storage.
10. Feldman, Michael. *IBM Demos Record-Breaking Parallel File System Performance.* http://www.hpcwire.com/hpcwire/2011-07-22/ibm_demos_record-breaking_parallel_file_system_performance.html
11. Shilov, Anton. *Samsung Shows Off Prototype of 4TB Hard Disk Drive.* http://www.xbitlabs.com/news/storage/display/20110308081634_Samsung_Shows_Off_Prototype_of_4TB_Hard_Disk_Drive.html
12. Altavilla, Dave. *A Terabyte For NotebooksL WD Scorpio Blue 1TB Drive.* http://hothardware.com/Reviews/1TB-WD-Scorpio-Blue-25-HD-QuickTake/
13. *HITACHI Deskstar 0S03230 3TB 5400 RPM 32MB Cache SATA 6.0Gb/s 3.5" Internal Hard Drive -Bare Drive.* http://www.newegg.com/Product/Product.aspx?Item=N82E16822145493
14. *IBM – LTO Ultrium 5 – 1.5 TB / 3 TB – storage media.* http://www.amazon.com/IBM-LTO-Ultrium-storage-media/dp/B003HKLHZC
15. *HITACHI Ultrastar 7k3000 HUA723030AKLA640 (OF12456) 3 TB 7200 RPM 64MB Cache SATA 6.0Gb/s 3.5"*

Internal Hard Drive -Bare Drive. http://www.newegg.com/Product/Product.aspx?Item=N82E16822145477

16. *SuperChassis 417E16-RJBOD1.* http://www.supermicro.com/products/chassis/4U/417/SC417E16-RJBOD1.cfm