HIERARCHICAL STORAGE MANAGEMENT AT THE NASA CENTER FOR COMPUTATIONAL SCIENCES: FROM UNITREE TO SAM-QFS

Ellen Salmon, Adina Tarshish, Nancy Palm
NASA Center for Computational Sciences (NCCS)
NASA Goddard Space Flight Center (GSFC), Code 931
Greenbelt, Maryland 20071
Tel: +1-301-286-7705
e-mail: Ellen.M.Salmon@nasa.gov

Sanjay Patel, Marty Saletta, Ed Vanderlan, Mike Rouch, Lisa Burns, Dr. Daniel Duffy
Computer Sciences Corporation, NCCS GSFC
Greenbelt, Maryland 20071
Tel: +1-301-286-3131
e-mail: sjpatel@calvin.gsfc.nasa.gov

Robert Caine, Randall Golay
Sun Microsystems, Inc.
7900 Westpark Drive
McLean, VA, 22102
Tel: +1-703-280-3952
e-mail: Robert.Caine@sun.com

Jeff Paffel, Nathan Schumann
Instrumental, Inc.
2748 East 82nd Street
Bloomington, MN 55425
Tel: +1-715-832-1499
e-mail: jpaffel@instrumental.com

Abstract
This paper presents the data management issues associated with a large center like the NCCS and how these issues are addressed. More specifically, the focus of this paper is on the recent transition from a legacy UniTree (Legato) system to a SAM-QFS (Sun) system. Therefore, this paper will describe the motivations, from both a hardware and software perspective, for migrating from one system to another. Coupled with the migration from UniTree into SAM-QFS, the complete mass storage environment was upgraded to provide high availability, redundancy, and enhanced performance. This paper will describe the resulting solution and lessons learned throughout the migration process.

1. Introduction

The Science Computing Branch of the Earth and Space Data Computing Division at the Goddard Space Flight Center (GSFC) manages and operates the NASA Center for

Computational Sciences (NCCS).[1] The NCCS is a shared center providing supercomputing services and petabyte-capacity data storage to a variety of user groups. Its mission is to enable Earth and space sciences research through computational modeling by providing its user community access to state of the art facilities in High Performance Computing (HPC), mass storage technologies, high-speed networking, and HPC computational science expertise.

The largest workloads currently being performed at the NCCS consist of Earth system and climate modeling, prediction, and data assimilation. Input data for these applications come from many sources, including ground and satellite stations. Both computer and sensor technology have grown dramatically within the last decade causing a boom in the amount of data generated by these types of sources.[2]

The major groups that comprise the NCCS user community include the following:

- *Global Modeling and Assimilation Office (GMAO)*: consists of both the Seasonal-to-Interannual Prediction Project (NSIPP) and the Data Assimilation Office (DAO), produces ensembles of simulations of near-term climate and creates research-quality assimilated global data sets from multiple satellites for climate analysis and observation planning.
- *Goddard Institute for Space Studies (GISS)*: produces climate studies focusing on timescales ranging from a decade to a century.
- *ESTO/Computational Technologies Project*: develops the Earth System Modeling Framework (ESMF).
- *Atmospheric Chemistry*: research teams investigating the evolution of the composition of the Earth's atmosphere and its impact on weather and climate.
- *Research and Analysis Group*: a large collection of smaller research efforts.

2. Data Management at the NCCS

With over 3 Teraflops of computational capacity, the research performed throughout the heterogeneous environment of the NCCS uses large amounts of existing data for new computational studies while generating large amounts of new data from the output of these studies. In general, the total data stored at the NCCS is growing at approximately 125 TB of data per year, which includes both primary and secondary copies of user data. As an example of this net growth of data, during FY03, a total of 207 TB of new data was stored while approximately 143 TB of data was deleted. This resulted in a net growth of 64 TB of single copy data or 128 TB when duplicated. Complementary, the number of files managed by the Mass Data Storage and Delivery System (MDSDS) has grown from 3.5 million in 1999 to more than 10 million in 2003.

Figure 1 shows the linear data growth as measured at the end of the fiscal year (month of September) for the past five years of only the legacy UniTree data. This trend is expected to increase dramatically in the next few years as the diverse mass storage facilities at the NCCS are consolidated and with increased utilization of the computational resources.

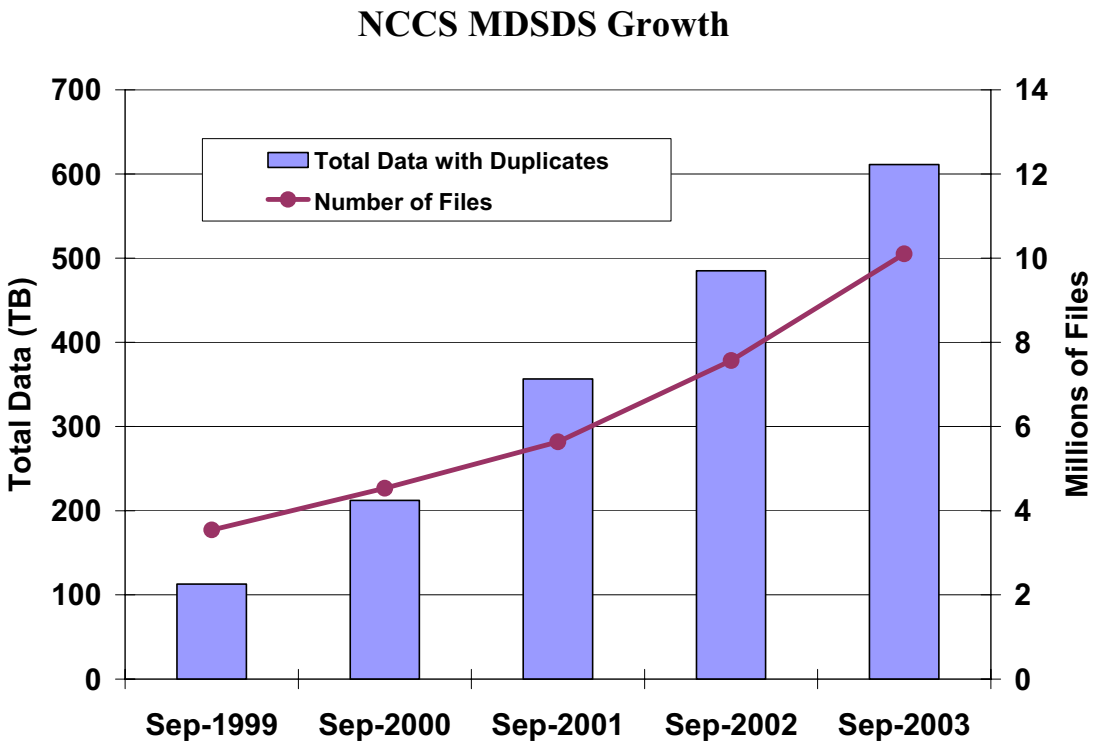

Figure 1: Growth of Mass Data Storage and Delivery System (MDSDS) data and files at the NCCS.

Throughout any given day, data is pulled from and stored to the MDSDS as jobs are run on the computational platforms. The NCCS measures the inbound and outbound data of the mass storage system and has seen traffic, files being transferred into and out of the mass storage system, of up to a total of 2.9 TB for a single day. Therefore, the resulting storage system must not only keep pace with the increase in overall storage, but also maintain the capability to serve larger amounts of data on demand.

3. Hierarchical Storage Management (HSM)

An HSM consists of different layers of storage capability for users to store and retrieve their data. Typically, a high speed disk cache is used as the first layer of storage and software is run to migrate files from high speed disk cache (1st layer storage) to slower tape media (2nd layer storage).

There are two existing HSM systems at the NCCS. The NCCS provides the MDSDS, previously running UniTree, for high-performance long-term storage for most NCCS user data.[3,4] A second system, which uses the SGI Data Migration Facility (DMF), supports the GMAO DAO users. This paper only discusses the replacement of the MDSDS's OTG's DiskXtender Storage Manager (DXSM) software, formerly known as UniTree Central File Manager (UCFM).

The UniTree management software runs on an aging Sun E10K with eight TB of Data Direct Network high performance disk storage. The MDSDS manages eight StorageTek (STK) Powderhorn 9310 robotic silos (five primary silos in the NCCS's primary building

and three secondary risk-mitigation silos in a building a mile away). For all user data, the MDSDS is configured to make a primary copy of files on tapes in the NCCS's primary building and a secondary copy on the tapes in the risk mitigation location.

Users access the MDSDS through the File Transfer Protocol (FTP) from any of the computational platforms or even their desktops. A home directory for each user is defined within a single MDSDS file system. Files that are put into UniTree are first copied into the MDSDS file system and then archived to tape and later released from disk cache according to NCCS policies. File retrieves are transparent whether the file resides on disk or must first be staged from tape; however, any retrieves of files from tapes incur a latency to load the tape into a tape drive, position the tape to the beginning of the file, and then copy the file to the disk.

While UniTree was a very reliable mass storage software system, by the middle of 2001, it became apparent that the recently modified capacity license cost model for UniTree was not compatible with the NCCS budget in light of the NCCS users' projected growth over the ensuing years. The NCCS began exploring alternatives and undertook a detailed feature comparison of four major storage management systems used for several years in high performance computing environments. The candidates were SGI's Data Migration Facility (DMF), IBM's High Performance Storage System (HPSS), Sun's SAM-QFS (also known as Sun StorEdge Performance and Utilization Suite), and UniTree. The various solutions were evaluated based on the following attributes:

- *Performance*: meet the needs for user requests for storage and retrieval of data.
- *Integrity/High Availability*: stable and safe environment more readily available than the existing HSM.
- *Flexible/Modular/Scalable*: allows for the maximum possible options for hardware and software and can scale with the users' requirements.
- *Balance*: avoid bottlenecks throughout the flow of data to the storage media.
- *Manageable*: tools provide a rich environment for administration and reporting.

4. Sun/SAM-QFS Solution

An internal panel evaluated the vendor responses and awarded the highest rating to the Sun SAM-QFS proposal. Notably, the Sun proposal scored high marks for its ability to be configured for high availability by sharing file systems in a clustered environment, its ability to "stream" the writing of tiny files to tape by combining them into "containers," and by having the largest customer base. Complementary to the Sun proposal, the NCCS also purchased more disk space and tape drive upgrades. The resulting system continued to leverage the existing investment in the STK hardware while providing a viable system to meet the future needs of the NCCS.

A Sun Fire 15K system was purchased and configured into two distinct domains. These two domains, along with multiple interfaces to each, provide a highly available system for the user community. Fully redundant, SAM-QFS provides the necessary storage management software to provide multiple file systems with storage, archive management,

and retrieval capabilities for a variety of storage media. The major components that make up the Sun SAM-QFS software are as follows:[6]

- *Archiver*: automatically copies online disk cache files to archive media. The archive media can consist of either online disks or removable media cartridges.
- *Releaser*: automatically maintains the file system's online disk cache at site-specified percentage usage thresholds by freeing disk blocks occupied by eligible archived files.
- *Stager*: restores file data to the disk cache. When a user or a process requests file data that has been released from disk cache, the stager automatically copies the file data back to the online disk cache.
- *Recycler*: clears archive volumes of expired archive copies and makes volumes available for reuse.

One of the key requirements from the outset of the transition to a new HSM was for the legacy UniTree data to remain transparently accessible to the user community through the new system. To facilitate the access of UniTree data on SAM-QFS, the entire file name space and directory structure of the UniTree system was recreated as directories and inodes in SAM-QFS. These inodes were basically placeholders, or links, to the original files in UniTree and contained an NCCS-defined volume serial number (VSN) and a "stranger" tape media type. Using the SAM migration toolkit, a set of libraries was created by Instrumental, Inc. to satisfy a stage request in SAM-QFS for a legacy UniTree file, which was identified to SAM by the "stranger" media type and the NCCS-defined VSN. Therefore, if a user requests a file that resides in UniTree, these libraries transparently retrieve the specified file from the UniTree system over a private network. Once the file has been retrieved from UniTree, it now exists within the SAM-QFS file system with two archive copies written to SAM tape and no longer needs to be retrieved from the legacy HSM.

Complementary to user driven access to the legacy data in UniTree, the NCCS has written Perl scripts to actively migrate the data from UniTree into SAM-QFS. These Perl scripts migrate files on a tape-by-tape basis and run "behind the scenes" to minimize the impact to the production environment. A single migration stream will secure files on a UniTree VSN from a well-defined list of UniTree tapes. This stream will get the current status of each file on that tape, i.e., whether or not the user has already migrated the file by retrieving it from tape or has even deleted the file. Next, the migration stream will begin to transfer the files over the private network using FTP. When the legacy files are retrieved to SAM-QFS disk cache, SAM writes two tape archive copies. After the migration stream has been completed, a separate analysis Perl script is run on each UniTree tape to verify that the files are in SAM-QFS. For quality control purposes, a checksum is run on every 100^{th} file. The current rate of migration is approximately 2 TB of data per day.

5. Conclusions

The integration effort of installing a new system to an existing High Performance Computing environment is difficult and requires much planning and effort. The

installation of the new Sun SAM-QFS system was no exception and many valuable lessons were learned.

- *Migration of Legacy Data*: The goal of migrating 100's of terabytes of data while still providing users with the ability to store and retrieve new files and transparently access legacy data is nontrivial and takes a significant number of resources. The amount of time and resources, i.e., tapes, tape drives, and network bandwidth, needs to be accurately estimated from the beginning and built into the integration plan such that users are not overly disrupted during the transition period as data is being migrated.
- *Test System*: It is important to have a test environment in which configuration modifications, such as operating system or storage software upgrades, may be tested without affecting the production environment.
- *User Account Management*: With two highly available domains on the new system, NCCS specific scripts were developed to synchronize user accounts between the two domains.
- *Pilot User Phase*: Before turning the system over to production computing, the internal NCCS staff and a set of pilot users were permitted access to the system. This phase allowed for a thorough testing of the environment before the full user community was allowed access.
- *Staff and User Training*: While the SAM-QFS system was designed to be as consistent as possible with UniTree, several training sessions were held with the staff and with the users to attempt to answer common questions. This allowed the user community to immediately begin using the new system and the staff to better support users from the beginning.
- *Software Upgrades*: Maintaining concurrency with the vendor's most recent release levels of operating systems and software is extremely important. Most vendors do not have the means to retroactively fix bugs for earlier release levels.
- *Security*: Define the necessary security requirements at the beginning of the process, and let those requirements drive the solution. It is more costly and disruptive to secure a system after it has been installed and patterns of use have developed by the user community.

The NCCS successfully transitioned the Sun SAM-QFS system into the production environment in September of 2003. The active migration of the more than 300 TB of data is slated to be completed in May of 2004. The new system has proven to be very reliable and capable of handling heavier loads than its predecessor. To date, the NCCS has seen tape activity, both user demand and migrations, exceed 9.8 TB for a single day.

As the NCCS continues to add computational capacity and as the user community continues to push the limits of modeling and assimilation to new heights, the HSM must evolve and adapt to the continued increase of requirements. The NCCS will incorporate the disk cache from UniTree into the production SAM-QFS system once the migration of the legacy data is complete. Also, the NCCS is analyzing the use of serial ATA, commodity based disk storage, as a second tier storage to sit between the high speed disk and slower tape. Finally, the NCCS is currently developing a data management system,

based on the Storage Resource Broker (SRB),[7] to provide user's with a single interface to storage and more control over their own data administration.

References

[1] http://nccs.nasa.gov.
[2] *Performance Management at an Earth Science Supercomputer Center*, Jim McGalliard and Dick Glassbrook.
[3] *Storage and Network Bandwidth Requirements Through the Year 2000 for the NASA Center for Computational Sciences,* Ellen Salmon, Proceedings of the fifth Goddard Conference on Mass Storage Systems and Technologies, (1996) pp. 273-286.
[4] *Mass Storage System Upgrades at the NASA Center for Computational Sciences,* A. Tarshish, E. Salmon, M. Macie, and M. Saletta, Proceedings of the Eight NASA Goddard Conference on Mass Storage Systems and Technologies, Seventh IEEE Symposium on Mass Storage Systems, (2000) pp. 325-334.
[4] *UniTree to SAM-QFS Project Plan*, Jeff Paffel, Instrumental, Inc., NCCS internal report.
[5] *UniTree to SAM-QFS Migration Procedure*, Daniel Duffy, Computer Sciences Corporation, NCCS internal report.
[6] *Sun SAM-FS and Sun SAM-QFS Storage and Archive Management Guide*, August 2002; *Sun QFS, Sun SAM-FS, and Sun SAM-QFS File System Administrator' s Guide.*
[7] http://www.npaci.edu/DICE/SRB/.