Mass Storage at NCSA: SGI DMF and HP UniTree

Michelle Butler, Robert Pennington and Jeffery A. Terstriep
National Center for Supercomputing Applications, University of Illinois, Urbana, IL

Abstract

The NCSA has been using UniTree since 1993 as the primary mass storage system for user's files and enterprise wide backups. The rapid evolution and growth of NCSA's production supercomputing environment, currently 512 Origin 2000 processors, a Power Challenge Array, and an HP/Convex Exemplar is putting increasingly critical requirements on the mass storage system. To meet this challenge, we have recently upgraded an HP/Convex Exemplar SPP-2000 running UniTree. In addition, we are evaluating SGI's Data Migration Facility (DMF) as a replacement for UniTree. This requires a transition from UniTree to DMF for our users and their data. (approximately 20TB). This paper discusses our observations of DMF and our future strategies.

Introduction

Mass storage is a critical component of a high-performance computing environment because it ensures long term continuity thus enabling researchers to conduct computational experiments that transcend restrictions on storage space and time available in a production supercomputing environment. In its ideal form, it should provide unlimited storage with infinite retention times. More practically, it must be able to retain thousands of times the data volume that can be generated by a computation for years and, most importantly, make it available in its original form upon request extremely quickly.

Conceptually, mass storage systems have remained virtually unchanged over the past decade. Computational users use *FTP* or similar mechanisms to send data files to storage that is comprised of a server with a large disk cache that is backed up to tape by software on the server. This has worked extremely well for environments that are limited to a single machine room executing applications in batch mode. The requirements for the National Technology Grid and the consequent technologies are reshaping this concept into one in which mass storage must be ubiquitously available, transparently accessed and of very high quality in terms of performance and reliability anywhere within the Grid.

This document reviews the evolution of the NCSA mass storage system from its inception, analyzes the current systems usage and performance, and discusses plans for future enhancements that will meet the growing and changing needs of the Alliance.

System Evolution, 1985-1998

NCSA's Mass Storage System is constantly evolving to meet the ever increasing demands for high speed archiving of large data sets generated by NCSA's supercomputers. It has grown in capacity and capability as technologies have developed and come available. NCSA continuously strives to incorporate leading edge mass storage technology into the system while maintaining a robust production supercomputing environment. NCSA has deployed several generations of hardware and software as requirements grow and change and technologies mature.

NCSA's first mass storage system was on an Amdahl 970 system running an MVS operating system and using CFS (common file system) for mass storage. The system held 2TB of data and was growing by 100GB per month at the end of the first 7 years. The only systems allowed to communicate with the CFS system were the Cray supercomputers through a hyper-channel link using special commands to get/put files to the archive. As we moved beyond the Cray vector processing supercomputers there grew a need for better performance, open access to all systems, and a Unix like interface. UniTree was a startup mass storage system that many sites were investigating at the time.

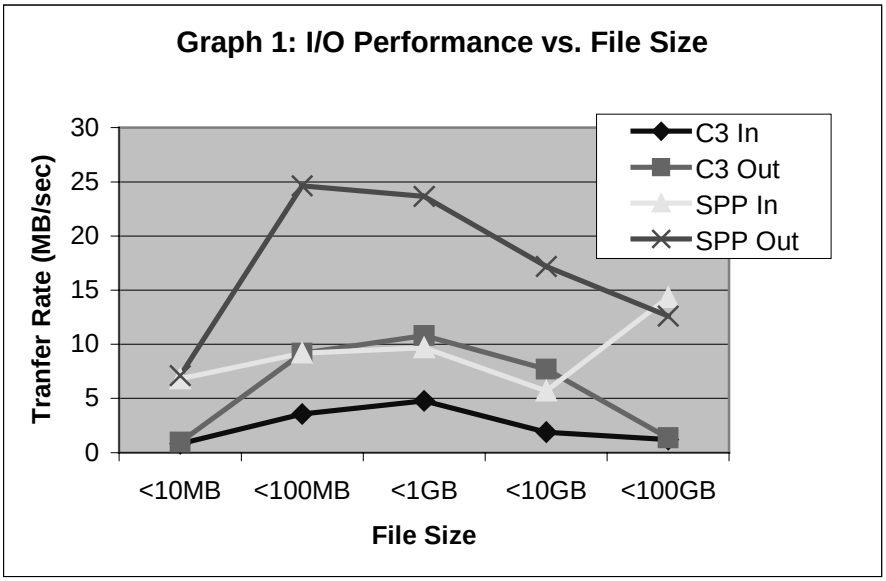
NCSA moved to a Convex C220 running Convex UniTree in June of 1993. The system had 256MB of memory and 100GB of disk cache. The CFS 3480 tapes were used in read-only mode and 8 Metrum tape drives were added to the Convex, each Metrum tape held 14 GBs and could sustain 1.5MB/s transfers. The archive had doubled to 4 TBs of data after only 2 years. Users had access to this system from any network connection and used a common FTP utility to manipulate their data in the archive. This corrected two of the largest complaints about CFS. As the use of the system grew, the C220 consistently ran out of memory for processes, and the disk cache was always full. Data could not be migrated to tape as fast as it was being archived.

In November of 1995 NCSA upgraded the Convex C220 to a Convex 3820 system with 2GB of memory and 300GB of disk cache. Eight new IBM Magstar drives were also installed over the next 3 years, each tape holds 10GB and the sustained transfer rate is 9 MB/sec. In addition, tape libraries for both Metrum and Magstar were added. The Metrum library holds 48 tapes while the IBM 3494 library for the Magstar holds 1200 tapes or 12TB.

In March of 1997 the mass storage system began to be used for system backups. This action increased the archive by 15TB in the course of 3 months. As more backups were added, a large strain was placed on the archive. The archive grew to over 30TB. While the C3 was handling the load, the system could no longer support any additional devices and had reached the end of its product life.

In January of 1998 NCSA migrated to UniTree+ 3.0 running on a HP/Convex SPP-2000. Because of the switched architecture, this system provides a significant increase in I/O performance and in the number of devices that can be attached to the system. The new system will allow for the expansion of devices and disk cache that the mass storage system depends on to keep up with the growing demands.

The system is currently equipped with 400GB of disk cache and 10 Magstar tape drives placed in either a STK Powderhorn library or an IBM 3494 library. The SPP2000 system is storing 2.5 TB of data per month of scientific data and enterprise wide. The system has been optimized to achieve the best possible transfer



rates over HIPPI, and is capable of storing and retrieving data at sustained aggregate rates of more than 40 MB/s.

Graph 1 shows the performance increase achieved when moving between the Convex C3 and HP/Convex SPP-2000. Stores to the C3 were limited to less than 5 MB/sec. The SPP-2000 achieved nearly 10MB/sec and almost 15 MB/sec on very large files. Retrieval rates were also enhanced and reached peak speeds of nearly 25 MB/sec.

High Performance System

Significant growth is forecast for the coming five years. The growth rate of scientific data sent to the mass storage system has averaged a 50% growth rate over the lifetime of NCSA and is currently at 2.5 TB/month, 60% new data, 40% rewritten. The projection is that the volume of scientific data created by Alliance users will grow to 200 TB by 2002 and the total archive will be 2-4 times this size with the presence of redundant copies and enterprise wide backups.

To meet the requirements for the coming years, the High Performance Data Management group at NCSA has begun to evaluate SGI DMF as near-term solution. In the longer term, we will be working with the Alliance Enabling Technologies Team C to identify and bring in-house critical new technologies in mass storage to support high performance computing applications across the Alliance distributed supercomputing environment.

NCSA has installed and is testing DMF on an SGI O2000 system. We are evaluating DMF as a possible replacement for UniTree for high-performance systems. DMF can provide enhanced performance, reliability and a clear integration path for mass storage and supercomputing on SGI systems.

SGI Data Management Facility

We initially tested DMF in three different configurations. The first was a traditional remote file system that was accessed using *FTP*. The file system was managed by DMF and tape drives were local to the server system. The DMF system was extremely flexible. Migration policies could be set based on uid, gid, filesize, etc. Files could be locked on disk, migrated to various types of tape drives, and duplicate copies could be created. The capabilities far exceed those of Unitree. Unfortunately, tape drives could not be shared between different data policies. SGI is developing a tape management facility (TMF) which should be available for beta testing in the next few months. Other weakness included the lack of a management GUI, lack of SNMP support, and system security is based solely on normal Unix means. To work within our environment Kerberos authentication would need to be added.

The second configuration that we tested was a DMF server running on a production 02K machine (client-side), managing a file system that used *FTP* to transfer the files to another DMF system (server-side) which managed a disk cache and tape drives. This configuration provides virtual disk, freeing the user from pro-actively archiving files using FTP. Data can be managed, whether local or archived, is treated just like local data. While this configuration is very powerful, but as multiple client-side DMF systems are added the data remains associated with file systems where it was originally written. Files can not be shared directly between DMF managed file systems.

This configuration had the archiving flexibility similar to the first, but added another layer of complexity to the architecture. The additional complexity requires system administrators to monitor multiple DMFs (databases, daemons, log files, etc) on different machines to locate problems with the environment. Again security was a concern, there is a hard coded clear-text "root" ftp password on the client-side DMF machine to allow connection on the server-side DMF machine.

The 3rd configuration we tested was a DMF server running and using HP/Convex UniTree as the background disk cache and tape server. This allowed us to have the flexibility of DMF for the file system and the tape management services of Unitree. It also provided a painless mechanism for introducing DMF into our computing facilities since data would not have to be immediately migrated from UniTree to DMF. As part of this we tested the *dmcapture* program that reads a NFS mounted partition (from UniTree) and builds a DMF file system. This worked out well after we turned off the UniTree+ sticky bit used to determine if the file was on tape or disk. (The program worked fine, but it kept the sticky bit, which meant nothing once the conversion to DMF, was complete.) Since our UniTree archive contains 2.5 million files/directories this process will take ~4 days.

For NCSA to use the DMF as a front-end of UniTree in order to migrate to DMF over time, there would need to be some changes in the way DMF finds the user's home directory. In UniTree, our users have home directories in different directory structures. There would need to be a "mapping" function added to DMF so that user's data could be found in this type of setup.

This architecture was similar to the second in that multiple mass storage systems are required. As this environment grows, and more compute servers are added, multiple DMF servers are needed on each compute server. As one can see, this is a very complex system that would not be easy to manage. Locating problems within the different systems is itself a daunting task.

In general, we have not found any problems with the installation or setup of DMF. Testing has just begun, we have not yet had to recover any databases or see how DMF handles bad tapes. We are very interested in DMF's ability to manage tape systems.

Summary

The NCSA mass storage system is an essential part of the supercomputing environment. It has undergone a number of evolutionary changes in the first decade and will evolve rapidly in the coming decade to meet the major challenges that accompany mass storage for the Alliance. Meeting these challenges entails a strategy of greatly enhancing the NCSA mass storage system for local supercomputing use and working with collaborators and vendors to extend these capabilities to the Alliance.

NCSA is also installing and testing DMF on an O2000 system to serve as a possible high-end replacement for UniTree. This DMF system could evolve into the mainstay mass storage platform for the SGI supercomputers and will provide enhanced performance, reliability and a clear integration path for mass storage and supercomputing on SGI systems.

The NCSA mass storage system is expected to ingest at least 200 TB of new scientific data over the next five years, with nearly the same amount of data being recycled by scientific users as part of ongoing computational efforts. Additional data storage requirements from joint data intensive efforts with academic and industrial partners will add measurably to this total, possibly on the order of an additional 25-50% of the data volume. Enterprise-wide backups for the supercomputing systems will bring the total on-site storage capacity needed within five years to approximately a petabyte.

Future work with SGI and the ET Team C in the areas of storage will play an essential role in the development and large-scale deployment of new technologies. The projection of this vision into the future requires a very substantial change in the way in which mass storage is viewed. The distinctions between file systems and mass storage systems and their association with computational systems must disappear, to be replaced by a large-scale computational system with a storage hierarchy, similar to a memory hierarchy. Work in some of the necessary areas is already being performed by researchers in the Alliance, including groups at UIUC, Argonne, and LCSE.