Storage Issues at NCSA: How to get file systems going wide and fast within and out of large scale Linux cluster systems

Michelle L. Butler
National Center for Supercomputing Applications (NCSA)
605 E. Springfield Ave
Champaign IL 61820
mbutler@ncsa.uiuc.edu
Tel: +1-217-244-4806
Fax: +1-217-244-1987

Abstract

This paper will discuss the history of storage at the National Center for Supercomputer Applications (NCSA) over the last fifteen years from inception to a four hundred terabyte archive. The paper discusses supercomputing requirements, hardware and software configurations, and the evolution of data management at NCSA. This paper also discusses the strengths and weaknesses of NCSA's different storage strategies, and gives a detailed discussion of the current system and how it is being evolved to meet the requirements of the TeraGrid computing systems, and large-scale Linux clusters.

1 Introduction

As NCSA, compute power has increased over the years, and so has the mass storage system to keep up with the ever-increasing rate at which data is produced. The NCSA mass storage system started in 1986 with thirty-six gigabytes of disk, a dual processor Amdahl performing twenty MIPS, with fifteen megabytes memory, and a single network adapter in the form of a 1.5 megabits Hyperchannel connection. The system has evolved to a single system configuration of sixteen 250MHz processors, twelve gigabytes of memory, three Hippi and six GigE network interfaces, and two terabytes of disk for overall I/O performance of two hundred megabytes per second.

2 History of Mass Storage at NCSA

In 1986, the first mass storage system at NCSA was an Amdahl running the Common File System (CFS) software package originally developed by LANL. This system was in production from 1986 to 1991 at NCSA, and served an evolving array of supercomputers from NCSA's original Cray XMP, to a Cray2, and a CRAY YMP. Access to mass storage was through a CFS client running on the Cray supercomputers. The data was staged to the Amdahl's disk cache, and then transferred through a proprietary protocol to the compute engine's disk. The only access to the mass storage system was through the Cray CFS client. Disk space on the Cray systems was purged after jobs completed, so users were responsible for storing files they wished to retain. The average file size was skewed by CFS's requirement to break data into chunks of two hundred megabytes. Files could not span tapes, and two hundred megabytes was the maximum that could be placed on the 3480-tape technology employed. All tapes were manually mounted, and redundant copies of every tape were made for off-site disaster recovery. Users began in later years to utilize other smaller data storage facilities. Direct access to their data was needed without mediation by an HSM, and then to a secondary machine like the Crays at NCSA.

The secondary staging was limiting, and the performance through the Hyperchannel was considered extremely slow for the times. User observed data rates were usually 1 Mb/s for a single stream, and multiple streams displayed a more dismal rate. New high-speed tape technologies were emerging, but the Amdahl could not be upgraded to handle those. The Amdahl was neither compatible with emerging tape and network technologies nor capable of advancing to follow on standard data protocols for data transfer.

NSL UniTree and UniTree from DISCOS were researched, and thought to be good products, but support in a 24/7 highly demanding production environment was questionable. Convex ported UniTree to their systems, and created a tuned version that was both faster and met NCSA's reliability requirements. NCSA wrote a conversion program for the move from CFS to Convex UniTree. The CFS databases were converted to UniTree format, and the system was "taught" how to read CFS tapes. Over 2 TB of data were converted, with a downtime of 3 days, to Convex UniTree. NCSA spent the next year rewriting all the CFS data tapes to the UniTree format, so code to read CFS tapes could be deleted at some future date.

2.1 Convex's version of UniTree

In 1991, NCSA moved to a C220I machine from Convex. The machine had dual processors and was wired for fast I/O. It had one hundred gigabytes of local SCSI disk, five hundred megabytes of memory, twelve 3480 tape drives manually mounted, and 1 Ethernet. The main user base still resided on the Cray2 and Cray XMP with a Convex 3880 machine coming into production as an additional compute server. The storage on the supercomputers was still purged as jobs finished, and users were required to store their own files and manage their own mass storage space. Accessibility was changed to a common FTP interface for all data, and data transfer performance improved because of the Ethernet interface(s). At first, the users liked the new procedures and were very happy with the FTP interface but, over time high-speed data networks were installed on the Crays, increasing network bandwidth, and mass storage transfers once again became a bottleneck. The data rate was too slow. User data rates were 6-8 Mbit/s (1MB/s). The one Ethernet interface could not keep pace with 2 systems running Hippi. Jobs were waiting on the Crays, and were wasting compute time in I/O wait states for the mass storage system to return.

The amount of data the system was ingesting was becoming more costly to store, and NCSA was forced to set storage quotas to limit users, mainly by encouraging them to improve their file management rather than by restricting the work they were able to accomplish. However, users reacted by storing their data in alternative, less reliable places that created more hardship for them. A new tape technology, Metrum 2150 tape drive, moved data at twice the speeds of the 3480's, stored seventy times as much on a tape (200 MB on a 3480 vs. 14 GB on Metrum), and a media cost was introduced to alleviate NCSA's storage cost problems. As data was written to tapes holding 14 GB/tape, the media expenditures of NCSA dropped dramatically. The Metrum tape drive specification stated drives should be used over 20% of the day. NCSA calculated that with 8 drives, that requirement could be met. NCSA also still dual-copied all data. The cost effectiveness of the Metrum tape medium enabled NCSA to lift user quotas. Over the next three years, additional Ethernet interfaces were added with increased disk cache allowing files to reside on disk longer. It became very apparent that a Hippi interface was needed to move data over the network faster, but the C220I machine could not be upgraded to include that interface. The Convex C3880 was being phased out as a compute server, and a large Thinking Machine CM5 was being brought into production. NCSA's mass storage system was "moved" to the C3880 machine. There was no conversion program needed. The C3880 had the same operating system and same hardware as the C220I machine. The databases were moved (FTP) to the new machine along with the tape drives. The data was purged from disk (all written to tape) on the C220I. When the C3880 came up, the data disks were empty, the databases showed all the data on tape, and six terabytes were "moved" to the new machine. All this took place during a normal downtime segment of less than 3 hours.

2.2 Continued Upgrades

The Convex C3880 machine (1994-1997) system was configured with eight Metrum tape drives, two gigabytes of memory, two hundred gigabytes of disk, eight processors, one Hippi interface, and two Ethernet interfaces. All traffic from the supercomputers was routed over the Hippi while traffic from other systems went over the Ethernets. This caused less congestion on the Hippi interfaces for slower data transfers. Users accessed mass storage through FTP and still managed their storage. During the production years of the C3880 archival storage machine, the CM5 was decommissioned, and SGI Power Challenge machines came into production. There was no longer one large machine, but several large machines all running jobs, and storing data. With many more machines capable of storing data through Hippi interfaces, a single Hippi interface could not keep up. Data streams started piling up with 3-4 concurrent transfers, driving down Hippi performance. The Hippi performance from the SGI's to the Convex was poor due to different revisions of hardware. The SGI PowerChallenge machines were capable at the time of 25MB/s, while the C3880 could transfer to the CM5 at 15MB/s, and only 3MB/s to the SGI machines. Tape technologies were also changing. The vendor was phasing out the Metrum tape. Therefore, new tape technologies were needed, but could not be connected on current machine. A new system was needed that could handle multiple Hippi interfaces (the latest revision), numerous simultaneous transfers and, as always, new tape technologies.

2.3 HP Exemplar X-class Machine

In 1997, NCSA purchased for the mass storage system server a HP X-class Exemplar machine. NCSA had stayed on the C3880 machine one year longer because there was not a strong I/O machine to move to until the Exemplar machine was ready for production. There was again very little conversion needed for the twenty-eight terabytes of data to be up and running quickly. The conversion was the same from the C220I to the C3880. All data was purged from disk, databases moved (FTP) showing all data on tape, old host turned off, devices moved, and new host booted with same old name. NCSA stayed on this machine for one and one-half years (1997-1998). This machine had eight processors, four gigabytes memory, five hundred gigabytes of SCSI hardware RAID disk, two Hippi interfaces and three Ethernet interfaces. Our user base started on the SGI Challenge and Power Challenge machines, and then migrated to the SGI Origin class machines. The 2

Hippi interfaces were divided up among the systems so that a "load sharing" could be achieved, giving users dual high speed data transfers into the machine. The new machine was capable of much more throughput than the C3880, so the simultaneous data streams count dropped dramatically. User scratch space was increased and more memory added to the production machines, but data management was handled as previously, an FTP interface for users to move/store data as jobs finished in batch queues.

The mass storage server system turned out to be a terrible environment. HP, who purchased Convex, phased out UniTree and Convex hardware support. Reliability of the system was questionable, it required a reboot every couple of days. NCSA did get some work done in spite of the problems by purchasing six IBM 3590 tape drives including NCSA's first tape robot, an IBM3494 library. NCSA copied all the Metrum data to IBM 3590 tape technology within one year because the vendor was phasing out the Metrum tape technology. The IBM3590 was faster than the Metrum, but did not hold as much data/tape. The IBM 3590 held at the time 10GB/tape. The cost difference was not significant enough to warrant changes in NCSA's storage policies.

The environment for the users remained the same. The aggregate throughput of the machine was much faster, but its instability drew many complaints. The Exemplar machine was able to stage/retrieve user data on both Hippi interfaces at 21MB/s (a combined total of 42MB/s). Normally there were 3 simultaneous transfers, but there have been as many as 12. The number of processors and machines in the Origin cluster continued to climb which in turn increased the need for more data streams to the mass storage system. Stability and aggregate throughput to keep up with the amount of I/O produced by our users were issues and NCSA again needed to upgrade

2.4 The switch to UniTree Inc and SGI

In 1999, NCSA evaluated HPSS, DMF and UniTree, Inc. storage systems. NCSA had a solid base in SGI's technology with much experience in the hardware and the software. UniTree, Inc. was selected to run on an SGI server. A new Origin eight-processor machine was purchased with four gigabytes of memory, two terabytes of locally attached fiber channel disk, three Hippi interfaces, and two Ethernet interfaces. UniTree, Inc provided a conversion program that rewrote the HP formatted databases on to the SGI in UniTree Inc's format, the data was purged from disk, devices moved. The capability to read HP formatted tapes was already in UniTree Inc's version. The new system came up with seventy-five terabytes of data on tape in a matter of hours. UniTree, Inc. on our SGI machine has proven to be reliable and efficient from its deployment in 1999 to today. The aggregate throughput of the mass storage system was 180 MB/s. During that time NCSA's user base was migrated from the one hundred and eighty SGI Power Challenge processors to fifteen hundred SGI Origin 2000 processors logically clustered into 10-15 individual machines. The user data rates were and still are 45MB/s for each stream across the Hippi network. .

The three Hippi interfaces on the mass storage system were load "shared" by dedicating a Hippi interface to the interactive machine, and splitting the traffic for the remaining Origin processors across the other two Hippi interfaces. The six 3590 drives were moved on to the new system, and a STK Powderhorn with seven 9840 drives and four 3590 drives was installed for a mixed media solution. This is the first time that NCSA has had a "mixed" media tape solution without decommissioning one of the two. NCSA used the 9840 tapes for the smaller files in the archive, taking advantage of the mid-load technology making time to first byte much faster. This small file threshold has changed over the years, but started out as 500MB or less. The 3590-tape technology was used for all other files, and all copy 1 data moved to an offsite facility. NCSA continued to run both IBM and STK libraries until the fall of 2001.

2.5 Upgrades to Origin 2000

Over the last two years, the mass storage system has grown in size and capability. NCSA started with eight 195 MHz processors, two gigabytes of memory, three Hippi network interfaces, and two Ethernet interfaces, an IBM library with capacity for 12 TB of storage, a Powderhorn library with capacity for 120 TB, ten 3590 tape drives, and seven 9840 tape drives. The system today has grown to sixteen 250MHz processors, with twelve gigabytes of memory, an ADIC AML/2 library with two sections for a capacity of 720 TB, an STK Powderhorn with capacity of 120 TB, six IBM LTO tape drives, ten 3590 tape drives, seven 9840 tape drives, eight GigE network interfaces, and three Hippi network interfaces. Its current throughput is 235MB/s with an archive size of 420 TB.

In the past two years, the user base machines have changed. NCSA now has fifteen hundred SGI origin processors with a mixture of 10 TB of disk. There are plans to deploy 15 TB more for production machines early in 2002. The mass storage system today supports a production IA-32 Linux cluster of 1024 processors and five terabytes of disk, a 180 node IA-64 (Itanium) dual processor Linux cluster, and an SGI Origin Array that will be phased out over the next two years as the Linux clusters move into production. The Hippi network will also be phased out, with GigE as the replacement. The performance study that NCSA has completed showed that the 45 MB/s single stream from the SGI's will *not* be matched, but the aggregate throughput of the GigE is greater because the handling of multiple concurrent streams is better. A single Hippi interface single stream runs at 45 MB/s and drops to 25MB/s for two streams, and 8 MB/s for three streams. A single GigE interface from SGI to SGI will transfer data at 25MB/s, and drops to 22 for two streams, and to 20MB/s for three stream. NCSA usually has 5-8 streams of data at all times.

The six TFLOP TeraGrid system will be the next big increment. The data that the mass storage system is ingesting is expected to continue to increase; however, predicting the growth rate and the necessary aggregate throughput needed has been difficult. Big jumps in CPU performance have inevitably produced more and more data, and the growth trends appear to advance along the same curve that is typical of other supercomputer centers. [1] If there is a big jump in CPU hours offered, the amount of data stored shows a proportional jump. But the network bandwidth into and out of the mass storage system that is necessary for applications is hard to predict. NCSA has been increasing aggregate bandwidth of the storage system after the need has been manifested.

NCSA has set a goal for 2002 of achieving 750 MB/s (three times current throughput) as the optimal performance for the mass storage system for the first year of the TeraGrid machine. The Itanium cluster is entering friendly user testing (March 2002). As 180 dual processor machines start storing data to the mass storage system through each systems' own GigE interface, observations will be gathered and adjustments will be made to local disk and archive systems as needed. Only time will tell if these predictions will ring true.

2.6 Hidden work for the mass storage system

The mass storage system at NCSA not only stores/retrieves user data, but also insures the integrity of the data trusted to the archive. In other words, if a file has been stored at NCSA's mass storage system, it will be retrieved. No files transferred properly to the mass storage system at NCSA have ever been "lost" or become irretrievable. There was, on one occasion, Hippi protocol inconsistencies between SGIs that contributed to a handful (<50) of files being corrupted before they reached mass storage. Those files were then retrievable, but still "corrupted". The duplicate copy has been a costly but wise investment. Media failures occur occasionally, but users at NCSA do not notice other than a file might take longer to retrieve than normal. NCSA is constantly rewriting data to new tape formats/media. Migrations in the past have been from the 3480 tapes to Metrum, Metrum to 3590, 3590 to 3590E or LTO, 9840 to 3590 or LTO. When purchasing a machine, NCSA has always included the background processes that need to take place to maintain the environment. Tape drives are not only needed for writing/reading of user data, but for repacking user data onto different tapes, possibly different tape types. The memory, disk cache, CPU, and tape infrastructure must be capable of handling these additional "hidden" tasks of a well-managed HSM.

3.0 Disk strategies for big iron

The large batch systems at NCSA serving supercomputing science over the years have changed quite a bit. Each increase in CPU capacity, memory, and new architectures has meant increased demands on the mass storage system. Sometimes, it has been more bandwidth into the machine for each stream, other times it has just been an increase in the amount of data stored. NCSA has benefited from other disk storage solutions that complement the mass storage system. Pools of local disk for the batch systems, and other smaller disk resources managed by the users for their own data have been highly effective. Each strategy tried has its niche for how it fits in the environment, but none of the solutions can do it all. Below are details on NCSA's file system strategies.

3.1 NFS

NFS has been used by every supercomputer that NCSA has placed in production. The Crays used it for cross mounting file systems to mount home directories and application software. NFS is slow. However it is easy, convenient, stable, compatible, and well understood by users. NFS is currently being used by NCSA for protecting the critical file systems of the large supercomputers. A failsafe server serves file space for user home directories as well as all application software. These file systems are exported from the failsafe system to the Origin Array, the Linux IA32 cluster and Linux IA-64 cluster. NFS is also used to cross mount all the local scratch file systems for each "type" of cluster. NFS is used by batch jobs to see all storage on the different batch machines, but users take a performance hit by using it for read/write operations.

3.2 Andrew File System

The Andrew File System (AFS) is heavily used more for the desktop infrastructure environment. NCSA hoped in 1994 that AFS would replace NFS for home directories and application software but the file system didn't have the performance required. AFS is used on the Origin cluster for a common link to center-wide installed software such as perl, email readers and the like. Some users do use AFS for data sharing to other environments at NCSA without FTP transfer, but performance is quite limited.

3.3 Local scratch

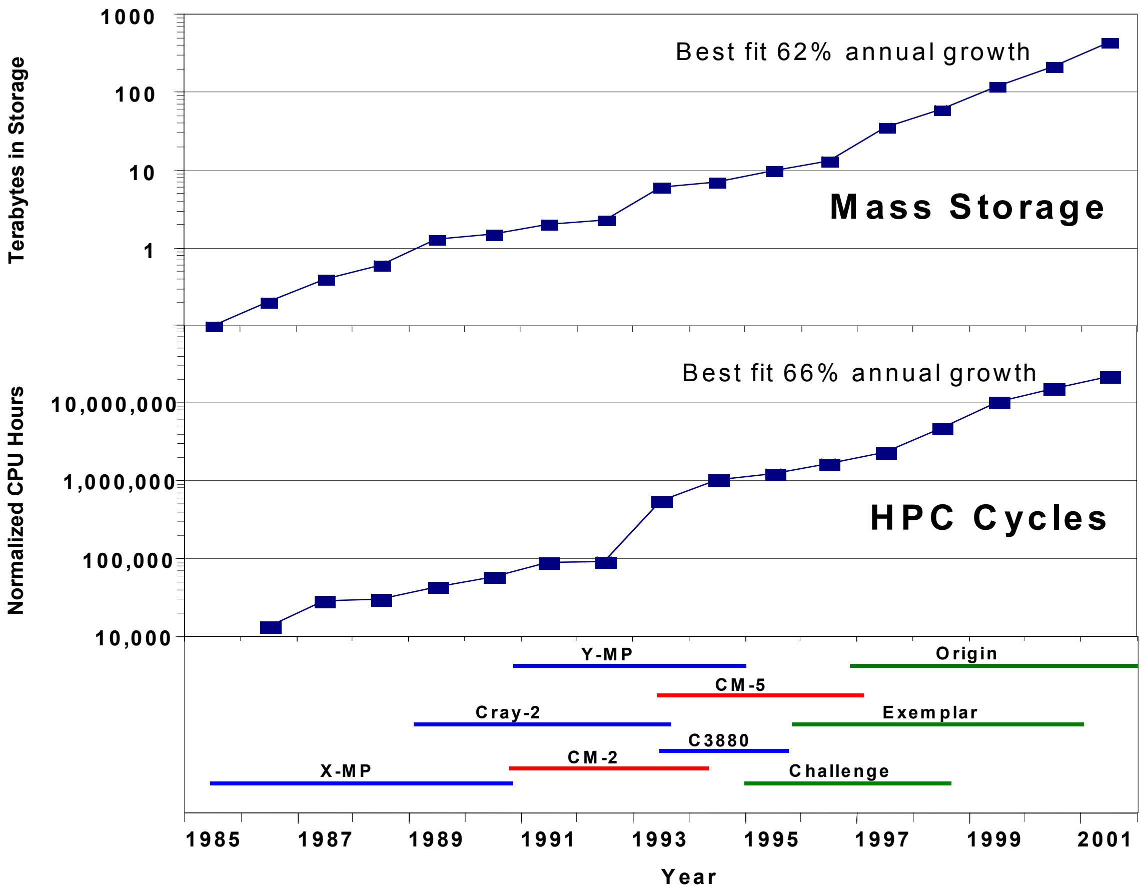
As described above, the large batch systems have local disk attached that is available to users for the duration of their batch job. As the jobs run, data may be retrieved from mass storage and before the job ends users are responsible to store their data back. NCSA has written a few "management" scripts for our users for doing persistent stores so that data will not be removed from scratch file systems until the files actually make it to the mass storage system. In the days of the Cray Super Computers users, had access to a gigabyte of disk storage for scratch space and that has grown steadily to where today NCSA supports file systems in the terabyte range.

3.4 Backup

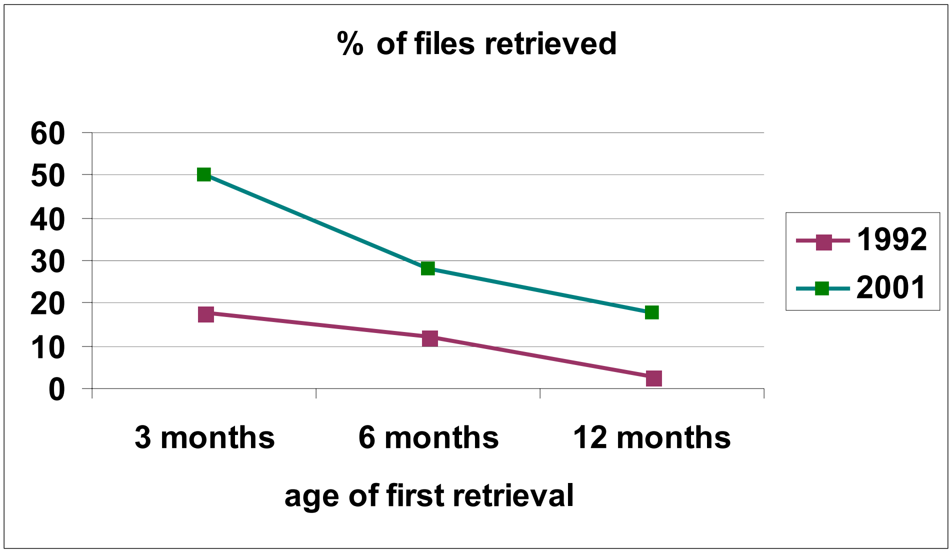
The backup system at NCSA also runs a UniTree storage system on a SUN 6500 machine. It has four IBM LTO tape drives, and shares the ADIC library with the mass storage system. This system handles one terabyte of data per week. NCSA backs up the AFS, NFS, /root, and /usr file systems for all the batch machines and all desktop machine/laptop/file servers. The data in the scratch file systems is too volatile and therefore are never backed up.

4. User and Storage patterns

The amount of storage at NCSA has continued to climb at a steady pace. Recently the growth has been more aggressive. The years 1997 – 2001 saw an 88% growth rate. As machine CPU hours continued to grow at close to exponential rate, the storage also followed faithfully. The chart below maps out the "normalized CPU hours" of the individual production machines at NCSA. The normalized hours have been calculated based on utilization of the machine, and then quantified to be equal among the different machine types. This allows us to equate cpu hours for all machines at the different supercomputer centers for NSF allocations of CPU hours. The bottom section of this chart shows the different machines that were in production during those years.

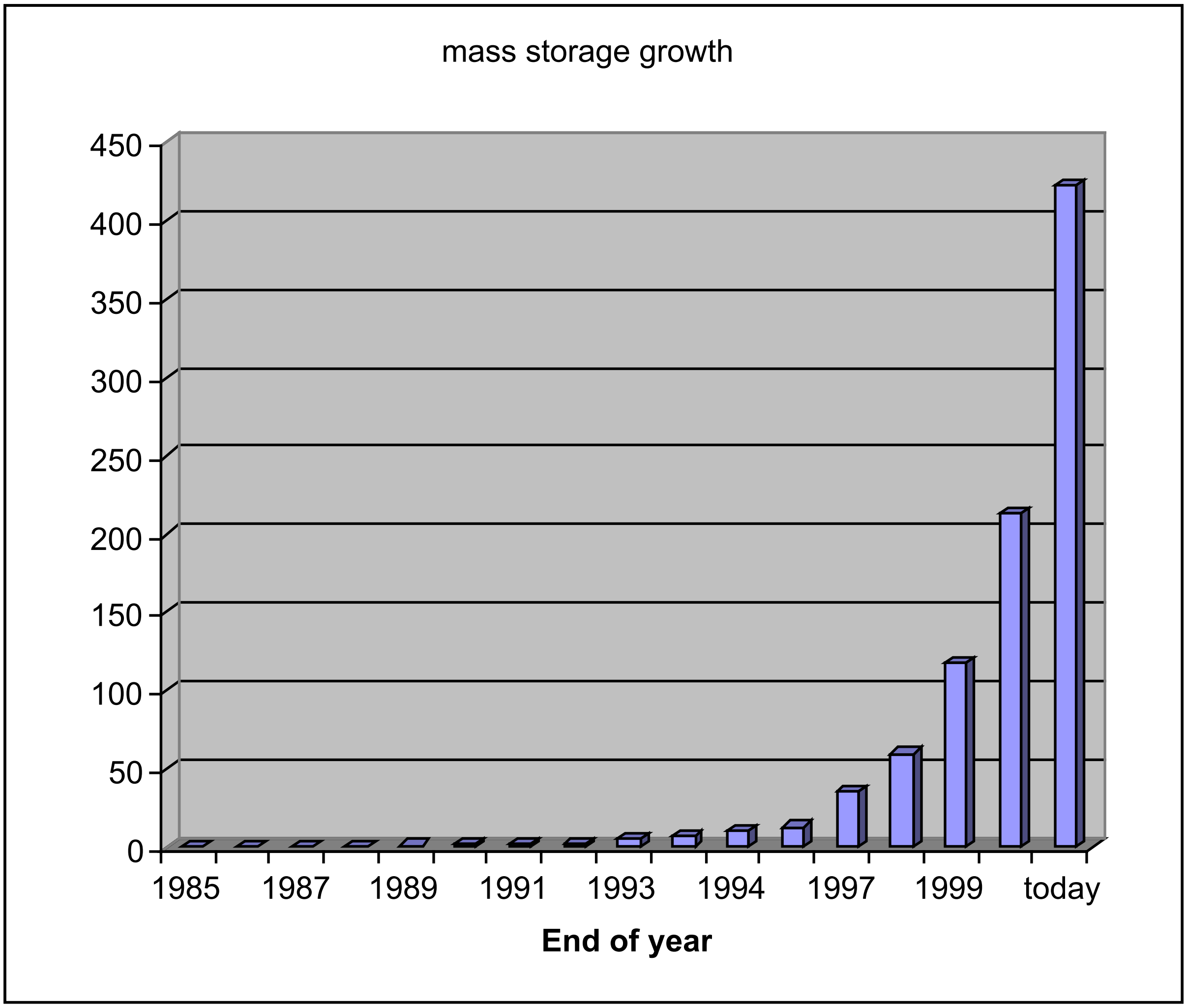


As the archive has grown, storage and retrieval patterns have changed. Large file archives historically have been read only [2] At the CFS conversion time, the size of the archive was 2 TB. UniTree was used primarily to store files that were never retrieved. The older the data, the lower the chance it would ever be recalled. Researchers try to predict what files will be used [3], but over the years, the "reuseability" of the files has changed dramatically. In 1992, as the graph below illustrates, 18% of files up to three months old were retrieved, at six months 12% were retrieved, and after 12 months 3% were recalled. Performance of the archive was unacceptable, and scientists found it faster to recompute data than to get the file from the archive. With increases in bandwidth and stability the data retrieval statistics have been changing, new files in the first three months in the archive have a retrieval hit rate of 50%, the first six months at 28% and drop only to 18% for data within its first year in the archive. So it is no longer a write only archive. Data storage performance was one of the most important criteria that the archive was judged on at NCSA, and now the increased speed and capacity have made data retrieval extremely important as well. Users are no longer recomputing, but retrieving data as needed, quite often, as the chart below shows. As scientific archives grow because of further research data derived from those archives, the role of data retrieval can only increase..

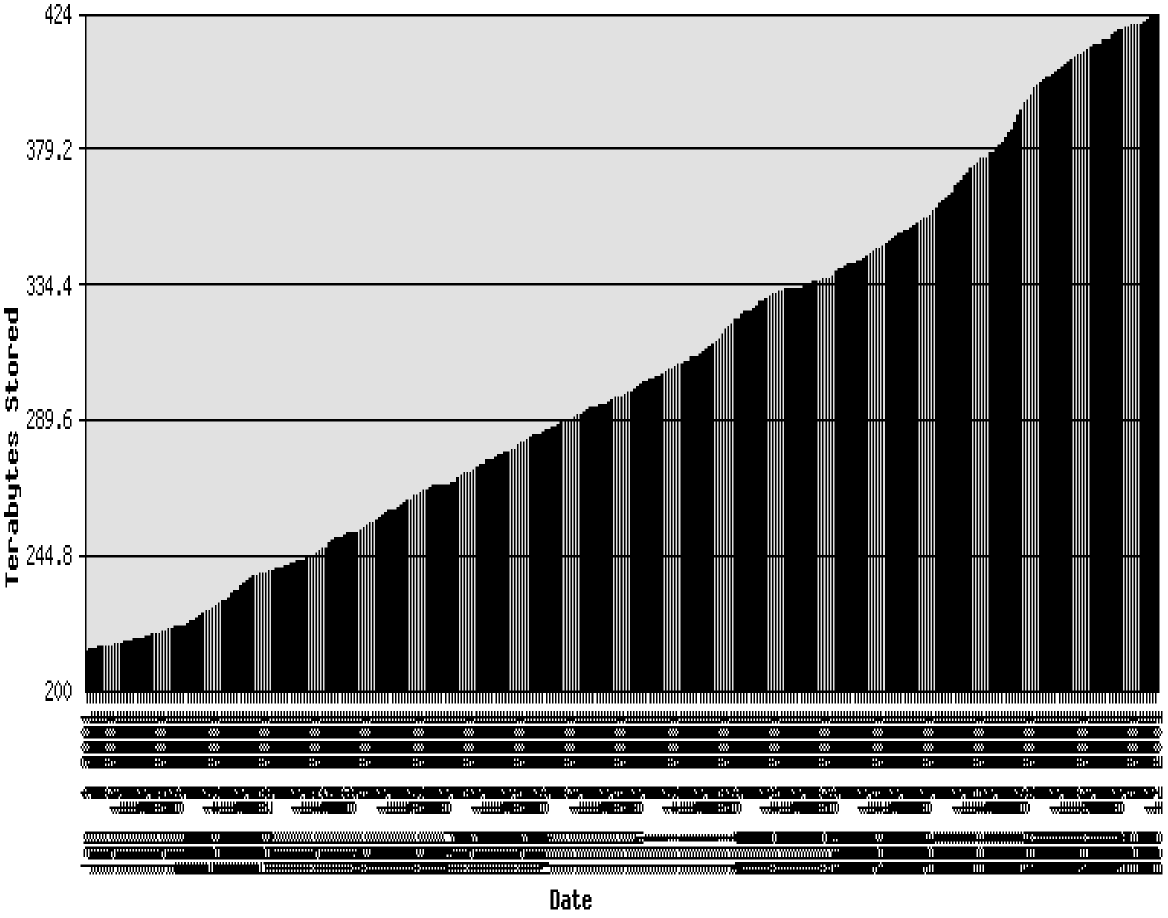


4.1 Growth for whole archive

Our growth patterns have remained much the same over the years. The archive size has been doubling about every year. The NCSA archive by this time next year will be close to a petabyte in size. Below is a graph of NCSA's overall growth. The first ten years are overshadowed on the graph by the huge amounts of data stored in the later years.



The graph of just year 2001 storage statistics for NCSA has a line for each day. The growth is very linear, and continues. For the TeraGrid, there will be a large increase in the data stored, but the amount is not known at this point. It is very hard to predict storage requirements for supercomputer centers [4]. As users have been given more resources in the past, they have produced more data, and storage seems to stay on the same curve as the normalized CPU hours of the machines.. The above graph does show a correlation to the CPU hours of a machine and the amount of data stored, but the number of CPU hours offered by a machine is not known. Within the next five years, there will be a technology switch again, as NCSA continues on the same curve; it is not known what is next for NCSA or supercomputing in general. [1]



4.3 Usage patterns and filesize

The average file size has also doubled in the last couple of years, but the average file size of our archive still seems small for a 400TB archive. Small files are normal for many large archives [4]. A chart of the average file sizes stored in the archive for the last six years shows that it has been increasing, but there are still very small files being used, while there are only a few files that are large (>500GB). This means that when purchasing drives and media types, the small files need to be considered. The small file is sometimes not brought into the mix when discussing mass storage, because large files are the norm, but as seen here, that is not true.

Year	Average File Size (MB)
1996	8.95
1997	13.75
1998	20.49
1999	38.97
2000	43.50
2001	68.88

The filesize growth may be attributed to increased capabilities of the processors so that transfers are no longer as time-consuming. The filesize certainly has not grown as expected, so maybe moving files that are 100GB or larger is still difficult, and a huge undertaking not only to stage, but to work with on the various production machines. As the average file size continues to grow, in 5 years NCSA users will be moving files > 100 GB with ease because of advances in data management and increased bandwidth.

Our top 10 users in FY 2000 stored:

	Files	TB
User1	4,391	3.2 (user 11 in 2001)
User2	259	2.8
User3	77,498	2.5 (user 9 in 2001)
User4	107,722	2.3 (user 1 in 2001)
User5	1,162	1.8
User6	2,743	1.7 (stays in slot 6 for 2001)
User7	3,790	1.6
User8	26,651	1.4
User9	8,757	1.3
User10	9,101	1.2

While in FY 2001 the top 10 users have stored:

	Files	TeraBytes
User1	328,394	9.4
User2	10,163	4.3
User3	23,404	4.0
User4	9,104	3.8
User5	1,871	2.5
User6	4,275	2.9
User7	2,427	2.1
User8	5,683	1.9
User9	30,033	1.9
User10	4,122	1.8

Just among our top ten users, the amount of data stored has considerably jumped. Our largest user in 2000 stored over 3 TB of data in 1 year. In 2001 our top four users each stored over 3 TB of data, with our top user in 2001 alone storing 9 TB. Another interesting point from the data above is that the top users at NCSA do not remain the same year after year. Only 4 users in the top 10 for year 2000 were in the top 11 of 2001.

4.4 Building for the TeraGrid machine

The NCSA mass storage system will be receiving another upgrade in Jan 2002 with an upgrade to six terabytes of disk. NCSA will also add an additional distributed disk server slated for production use in spring 2002. The second disk server will be an SGI Origin 3200 with four processors and two gigabytes of memory. The 3200 machine will have six terabytes of disk and ten GigE interfaces for a throughput of 250MB/s. NCSA is researching currently how to split data across the machines, with criteria based on uid, gid, original IP address, or file size being investigated. The new system combined with the current system makes the disk cache twelve terabytes with a real aggregate throughput of 450MB/s. NCSA will be also adding ten more IBM LTO tape drives. In late 2002 a 3rd distributed disk machine, an SGI Origin 3400 with aggregate performance of 300MB/s is to be put into production. This will bring the aggregate mass storage throughput to our goal of 750 MB/s. This goal has been based on the TeraGrid machine's predicted performance and the cost analysis of additional bandwidth/throughput for the mass storage system.

Now that NCSA has machines that can handle data at very high rates, and grid and user portal environments are being deployed, improved user tools are needed to move data from place to place. Some important deficiencies relate to inadequate descriptions of what data are available, where the data are located, and how and under which condition users may access the data [5]. The tools that NCSA has given our users have not changed from some form of FTP. NCSA is working on porting GRIDFTP from the Globus group onto the UniTree server so that the FTP transfers will be in parallel to the mass storage system. These tools are also being added to the distributed parallel file systems as explained below. We are incorporating GRID data technologies and working with the Globus group [6] at ANL to enable a grid environment of data being moved, replicated, and archived for all grid users. Gridware from Globus will help users take advantage of different data storage components with in the Grid, and aid the users in data management issues.

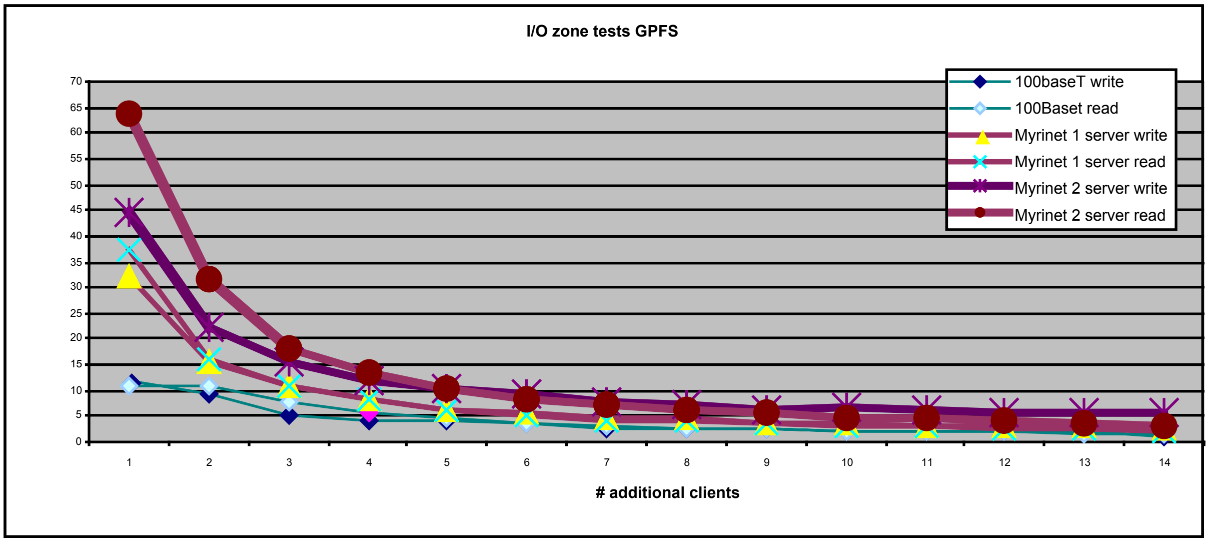
5.0. Linux Clusters Storage

NCSA is looking at many different file systems that might be able to accomplish our goals for the TeraGrid machine, and one standout is the Global Parallel File System (GPFS) from IBM. This is the linux port of GPFS to IA32 architectures from the SP2 machines. GPFS has been running at NCSA since October 2001. GPFS has three major components: a) the disk server is the machine with the disks attached; b) the GPFS server is the metadata server; and c) the individual client. A GPFS file system client must be installed on each system. Each system can then see all the data. GPFS can scale up by adding more servers and clients. GPFS can have multiple servers hosting the *same* file system or individual file systems as needed. NCSA has tested up to 120 clients and 8 servers all seeing the same single file system. GPFS has high availability options so that there is fail-over for disk servers and GPFS servers. Users interface with the native I/O commands to the file system, and all clients can read/write to the same file system and even the same file. Files are distributed across multiple servers by GPFS so that one user can gain access to the entire GPFS file system with all servers writing data at once. The performance does decrease as expected as more I/O requests are added from there.

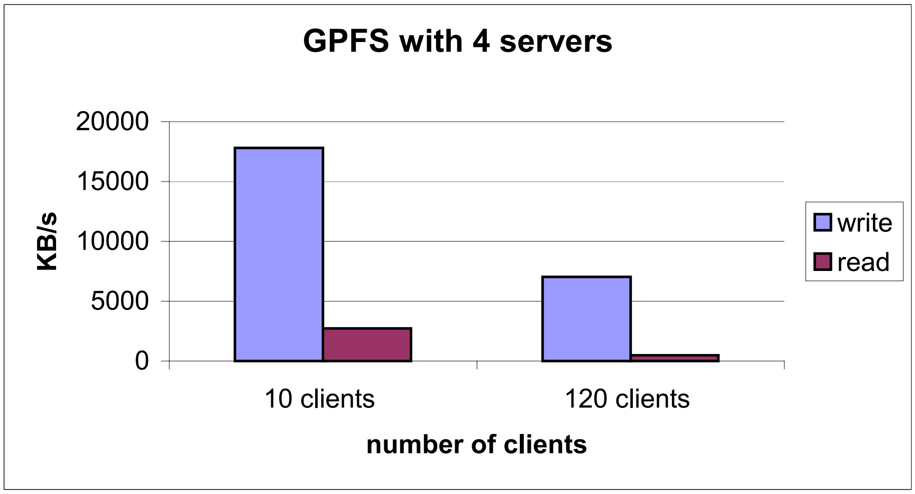
NCSA thinks that this is a very strong product with a very good team behind it. GPFS relies on a very fast low latency network for good performance to be observed. Since the changes in Myrinet driver in release 1.5, GPFS made great strides in reliability. GPFS is a file system for a single system only, there is no data sharing with other systems. A follow-on phase of GPFS development with IBM is a mixed GPFS cluster file system.

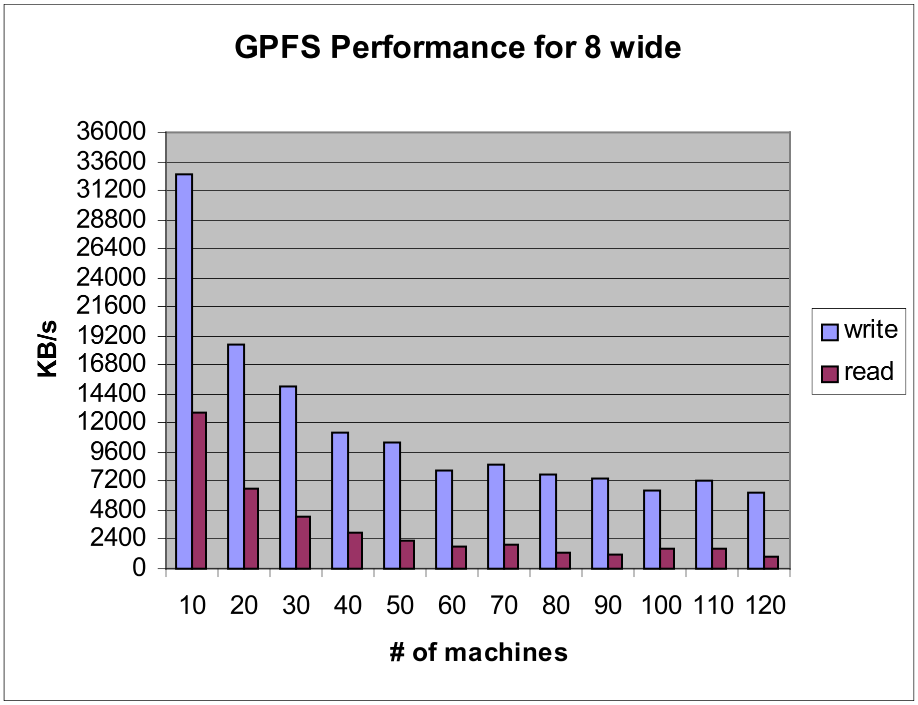
The mixture would be IA64 and IA32 clients and servers for a single GPFS file system. NCSA wants to add the Globus toolkit to GPFS, so that parallel data transfers can be used to move data out of the linux cluster machine to other grid systems or a mass storage.

The chart below compares the performance NCSA had with Ethernet and Myrinet. Myrinet has the best performance. The chart also shows the performance of 2 servers running on Myrinet. The performance that one client receives shows that the single client can gain the entire GPFS file system pipe. The performance scales down from there. These runs on the file system were done before several updated releases of the RedHat kernel with significant I/O changes.



The performance achieved running 4 and 8 servers and various numbers of clients is shown in the next chart. The 4 wide servers numbers were run before tunables for the kernel were made. The 8 wide tests have the kernel mods, but the SAN disks haven't been tuned yet. All clients write a 256 MB file simultaneously**.** Neither IBM nor NCSA is satisfied with the performance, and both are working on that part of this project. Problems are thought to be in the 7.1 kernel. Reads for a 10 wide test of GPFS are >12 MB/s on average, and > 31MB/s for writes.





6.0 Conclusions

The mass storage system at NCSA has evolved over the years. It started out as a small system with a slow interconnect and evolved to a very large system with many fast network interfaces. The supercomputer machines providing the bulk of the data to the mass storage system have also evolved. The machines started out as one system with a few CPU's, changing to a few systems with many CPU's, to many machines with few CPU's. File systems on the supercomputers have also changed, but users must do their own data management. They decide where to put their data depending on their applications. The interfaces for users to move data are still the rudimentary FTP tools. NCSA is making great strides to incorporate Globus grid tools into clients and servers for utilization of parallel data transfers, and better data management.

NCSA is adding a distributed data cache machine to its mass storage architecture to enable more simultaneous data transfers as the TeraGrid machine is built. More data cache machines will be added depending on how much aggregate data throughput is needed. History has shown that NCSA's data archive is growing at almost the same rate as the normalized CPU hours on the production machines. This is not hard to predict for maybe a year out, but gets harder the farther out one goes. The throughput is the hardest question. Not only do the mass storage archives need to keep up with the production machines on the LAN, but also as GRIDs gain users the amount of data coming in/out from production machines on the WAN will become an issue.

NCSA is looking at many different file systems to provide the best environment for our users. GPFS from IBM is being tested and beginning a friendly user period at NCSA. However more needs to be done to "share" data between these individual compute islands. Moving the data to the machine an application is running on as needed is a step in the right direction, but more needs to be done in this arena. Most of these tools today also deal only in flat files while databases are gaining respect and speed in the supercomputing environments.

References

1. Horst D. Simon, William T.C. Kramer, and Robert F. Lucas, "Building the Teraflops/Petabytes Production Supercomputing Center" *EuroPar* '99 in Toulouse, France, September 1999

2. Heinz Stockinger, Kurt Stockinger, Erich Schikuta, Ian Willers. "Towards a Cost Model for Distributed and Replicated Data Stores". *9th Euromicro Workshop on Parallel and Distributed Processing PDP 2001,* Mantova, Italy, February 2001, IEEE Computer Society Press

3. Timothy Gibson and Ethan Miller, " An Improved Long Term File Usage Prediction Algorithm," *Annual International Conference on Computer Measurement and Performance (CMG '99)*, Reno, NV, December 1999

4. Joshua C Neil, "Characterizing Long Term Usage of a Mass Storage System At a Super Computer Site", *Eighteenth IEEE Symposium on Mass Storage Systems* IEEE 2001

5. CODATA Committee on Data for Science and Technology, Working Group on Archiving Scientific Data, http://www.nrf.ac.za/codata/

6. Ian Foster, Steve Tueke, Carl Kessleman, http://www.globus.org