LEGIBILITY NOTICE

A major purpose of the Technical Information Center is to provide the broadest dissemination possible of information contained in DOE's Research and Development Reports to business, industry, the academic community, and federal, state and local governments.

Although a small portion of this report is not reproducible, it is being made available to expedite the availability of information on the research discussed herein.

1

LA-UR -88-2752

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

LA-UR--88-2752

DE89 000393

TITLE PROFILES IN MASS STORAGE: A TALE OF TWO SYSTEMS

AUTHOR(S) Maurice W. Collins
Marjorie Devaney
David Kitts

SUBMITTED TO Supercomputing '88
Orlando, Florida
November 14-18, 1988

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

[illegible]

Profiles in Mass Storage: A Tale of Two Systems

Bill Collins
Marjorie Devaney
Computing and Communications Division
Los Alamos National Laboratory
Los Alamos, NM 87545

David Kitts
Scientific Computing Division
National Center for Atmospheric Research
P. O. Box 3000
Boulder, CO 80307

Abstract

The Los Alamos Common File System (CFS) and the NCAR Mass Storage System (MSS) are file storage and file management systems that serve heterogeneous computing networks of supercomputers, general purpose computers, scientific workstations and personal computers. This paper details philosophical, implementation and performance aspects of the two mass storage systems. Areas covered include the computing environment, the user interface, storage strategies and file movement strategies.

Environment

Both the Los Alamos and NCAR computing environments are dominated by the use of supercomputers. Supercomputers are used at Los Alamos to solve a wide variety of scientific and engineering problems while at NCAR the use is focused on research in the atmospheric and oceanographic sciences. As would be expected of a large supercomputer site, both have large data storage and data retrieval requirements. Both systems currently store over seven terabytes of data, are growing at the rate of over two terabytes per year and transfer over 50 gigabytes of data each day. Although Los Alamos has significantly more computational power and users, the NCAR data storage requirements are equivalent due to their more data intensive applications and large data bases.

The major uses of a centralized file storage system at Los Alamos and NCAR are:

- Archival Storage. Large quantities of infrequently accessed data are saved for long periods of time.

- Inactive Files. Files are off-loaded from the expensive supercomputer disk during periods of time when they are not being used.

- File Backup. Users, operating systems, and servers throughout the network have a reliable place to back up their permanent files.

- File sharing. Files may be conveniently shared between users and machines in a heterogeneous network.

The CFS Environment

The Los Alamos Integrated Computing Network (ICN) is a large scientific computing network that provides services for over 8,400 users. The ICN is a network of many different machines running eight different operating systems (CTSS, UNICOS, UNIX, VMS, NOS, MVS, VM/CMS and MS-DOS). The machines include supercomputers, general purpose computers, scientific workstations and personal computers. These resources are summarized in Figure 1. Network support servers are provided for such functions as file storage, output processing, data import/export, access control, accounting, and batch job submittal/control The ICN is partitioned to allow classified computing, privacy computing and unclassified computing. The Common File System (CFS) provides a centralized file storage and file access capability for all machines and servers in all partitions [1]. The success of centralized mass storage systems has resulted in the CFS software being installed by seventeen other computing sites.

Worker Machines	QTY
CRAY X-MP/416	2
CRAY X-MP/48	2
CRAY X-MP/24	1
CRAY 1	4
IBM 3090-200	1
CDC CYBER	4
DEC Distributed Processors	160
SUN Workstations	400
IBM Workstations	100
Other Workstations	75
CRAY Y-MP/832 (4th Qtr. 1988)	2

Los Alamos Computing
Network Resources

Figure 1

At Los Alamos, the supercomputers are primarily used interactively during the day, and for the running of batch production jobs during the night. The interactive use includes program development, job setup, running short jobs, and output analysis. Users submit production jobs to a job control server which selects where and when to run each job, using job and input files stored on CFS. The production jobs generate restart files and graphics files which are stored on CFS as they are generated. During the day, users analyze these graphics files using supercomputers and high performance graphics terminals.

Distributed processors have been used for a number of years to provide a general purpose computing capability at remote sites and to provide special computing services for the entire network. More recently, scientific workstations and personal computers have been connected to the network to provide stand-alone computing capabilities and pre/post processing for supercomputer applications. The distributed processors, scientific workstations and personal computers have the same access to the CFS capabilities as do the supercomputers. Each machine has a user-level utility for request/response communication with CFS, and has implemented the file transport protocol to transmit files with CFS and any other machine or server in the network.

The NCAR MSS Environment

The NCAR Scientific Computing Division provides supercomputing resources and services that support research in the atmospheric, oceanographic and related sciences. These functions are supported by computers and operating systems from various manufacturers that include COS, MVS, VM/CMS and VMS, and in the future, the UNICOS operating system will also be supported. These operating systems are run on supercomputers, general purpose computers, and workstations. Their resources are serviced by the NCAR Mainframe and Server network (MASnet) and TCP/IP networks. These networks supply communications for file storage, import/export of data, access control, and batch job submission. These resources are listed in Figure 2. The NCAR MSS provides centralized file storage and access capability for all machines and servers in this network and a high-speed data path that is a unique feature of this system.

At NCAR, supercomputers are currently operated in a batch job mode. Interactive use is available but lightly used. Future plans to support UNICOS as the operating system for supercomputers will facilitate user interactive access. Currently, a user selects the supercomputer to be used for execution of a job and submits a batch job control program via the MASnet system. Input files stored on the MSS are accessed as needed by the program. Using this facility, large production jobs are restarted and massive amounts of data acquired and disposed to the MSS. Supercomputers at NCAR have limited local storage in relation to the amount of data processed in some sessions. This limitation forces users to save data and restart-images on the MSS, because space allocation mechanisms used at NCAR will delete files on the supercomputer that are not in use, in order to make room for current space demands.

Worker Machines	**QTY**
CRAY X-MP/48	1
CRAY-1A	1
IBM 4381-P14	1
SUN Workstations	20
Other Workstations	50

NCAR Computing Resources

Figure 2

Implementation

Both the NCAR MSS and CFS software have been implemented to run as application programs under the IBM MVS operating system without any changes to MVS. MVS and the System 370 architecture provide software and hardware interfaces for high performance commercial storage systems, efficient I/O processing, large memory addressing, a rich software environment upon which to base a file storage system, and a wide range of processors to run the software. PL/1 was chosen as the implementation language by both sites since it provides a good multitasking environment and requires only a minimum of assembler code. The software of each system has elements providing the functions of the IEEE Mass Storage Reference Model [2]. These elements include the Name Server, Bitfile Server, Storage Server, Physical Volume Repository, and Bitfile Mover.

Implementation of the NCAR MSS

The selection of MVS/XA was done after examining operating systems of several manufacturers and was totally independent of the Los Alamos selection. Each of the processes that comprise the MSS constitute a separate MVS batch job. These batch jobs run independently of each other and communicate using the NCAR Inter Process Communication Subsystem (IPCS). The IPCS is an independent process which passes "messages" between other processes. The processes may be separate MVS batch jobs or separate tasks within the same batch job. IPCS is not a protocol. The protocol is defined by the two unrelated communicating processes. IPCS is only a message passer. An important feature of IPCS is the ability to add new processes to the MSS system without modifying all of the MSS software and, in many cases, without modifying any of the MSS software. New processes can be easily integrated into the MSS system and test processes can be executed concurrently with the production system.

Utilizing individual batch jobs enhances the MSS system availability and reliability. The use of the IPCS has isolated software failure. Modules that generate a problem can be automatically restarted avoiding failure in associated processes. If the MSS were run as a single batch job, a software failure could interrupt the entire system. Enhanced and maintenance versions of processes can be installed while the system is running by simply restarting the appropriate batch job.

IBM channels allow device independence that NCAR was unable to obtain using other products. This flexibility is demonstrated by the multiporting of the IBM storage devices between the IBM 4381 control processor and supercomputers. This access of storage devices is a unique implementation method which allows a supercomputer direct communication to IBM 3380 disk and IBM 3480 tape cartridges. This data link was discussed in another paper [3], and in this paper it is referred to as the Local Data Network (LDN).

IBM access methods were not used as an interface to the 3380 disk farm and the 3480 archive storage devices. Common and individual problems were encountered on each device which prevented the use of vendor access methods. A large physical record size (40 KB) facilitated data transfer over the LDN. However, physical records of this size are not supported under MVS/XA.

Implementation of CFS

CFS has progressed, with minimal changes, from the original use of 3350 disk, the 3850 Mass Storage System and a 370/148 processor to the current use of 3380 disk, 3480 tape, a 3090 primary processor and a 4381 backup processor. In large part, this was due to the extensive use made of the MVS software. This use includes the access methods for storage system input/output, standard IBM tape labeling, and the use of VSAM data sets for the CFS directory system. However, key areas of CFS such as file cataloging, space management, volume management, volume mounting and tape drive selection are done with CFS software to satisfy the CFS performance, reliability and security requirements, and so that changes and enhancements can be quickly made to deal with new performance, operational and user requirements. The CFS software is very robust, and is able to intercept and recover from most hardware errors without impacting system availability.

The CFS software consists of the Production Program which provides the File Server, Name Server, Storage Server and File Mover functions of the Reference Model, the File Migration Program which decides where in the storage hierarchy each file should reside, and support programs that perform system management functions such as directory backup, directory recovery, volume labeling, tape analysis, performance measurement, etc. The implementation of the Production Program is based upon the primary or driving functions being the user requests (get a file, store a file, etc.). Network input is done by a single router task which allocates a request task for each new user request and then directs subsequent input for the request to that task. Except for network input, the request task performs all processing of the request including storage system input/output, directory read/writes and network output. All the program modules are reentrant so they can be used simultaneously by multiple tasks. This architecture has resulted in a very efficient and easy-to-modify system. The use of external servers such as a storage system on another machine is not precluded.

User Interface

CFS and the NCAR MSS are complete file management systems which provide file storage, file access and file management capabilities to a diversity of users that include interactive humans, batch programs, operating systems and other network servers. The file management includes a human oriented naming capability as contrasted with the machine oriented "bitfile ID" of the Reference Model and a tree structured directory system that allows users to organize the storage of files in a manner that facilitates the file access and file management. Users have a device independent interface to CFS and MSS where the systems determine where files will reside based on file size and file activity. CFS and MSS do not support or recognize any data structure. Files are stored as bit strings. When the user wishes to access data in a file, the complete file is transmitted to the machine. If a user changes the file, the complete file must be transmitted back to the storage system. CFS and MSS provide the capability for users to group files on tape volumes so the files may be efficiently retrieved as a "family."

Both systems have kept the user interface at a very simple level by keeping most of the decision process at the user level. The user must take explicit action to store, retrieve, delete, convert and backup files. This requires the user to be more knowledgeable and to do more work, but gives more control and flexibility.

CFS User Interface

The CFS user interface provides requests to create, delete, modify, move, merge and list directory structures and directory information, and to save, replace, get, delete, move and copy files. An application level utility, called the CFS Interface, is available on all network machines to pass text string requests from the user to CFS. CFS will parse and execute the request, and send a text response back to the Interface utility which will pass the response back to the user. The Interface utility is easy to implement and easy to maintain since existing CFS requests can be modified and new CFS requests added without affecting the Interface utility. The Interface utility presents the same interface to users on all of the many operating systems for which it has been implemented. Other utilities that use the Interface utility provide higher level interfaces that are often tailored to specific operating systems and environments. Examples are a graphics window interface for UNIX based workstations, a menu driven interface for MVS, program callable interfaces for different languages and various utilities that provide wildcard capabilities.

The NCAR MSS User Interface

The user interface is determined by the operating system used for communicating with the NCAR MSS. Almost every computer at NCAR has existing software to transfer data between itself ard others via the MASnet. Many of the operating systems have built-in facilities to do this, such as the COS ACQUIRE/DISPOSE mechanism[4]. The NCAR MSS uses native operating system interfaces, thereby reducing the need for users to learn a new interface to read and write files on the MSS.

When a user generates a new file which has a tree structure of undefined directories it is not necessary, and no capability is provided, to create empty directories. All necessary directory structures are created when a file is created. When the last entry in a directory is removed, the directory is deleted from the tree structure. In addition, if any directory parent becomes empty it is also removed. Separate utilities are provided to modify directory and file names, touch a file to reset retention time for archival purposes, delete files, and to request status and all attributes of a file or directory.

Storage Stragegies

Storage strategies used by the NCAR MSS and CFS are very similar. They both use IBM 3380 disk for active file storage and group the disk files by size to minimize fragmentation problems. Migration algorithms move inactive files from disk to IBM 3480 archival tape cartridges based on idle time and file size, and archival files are migrated back to disk when they become active.

Neither system allows a file to span multiple volumes and thus restrict the maximum file size to about 200 megabytes. Each system has its own unique format for files. Therefore, both NCAR and Los Alamos have special systems to "import/export" external files. Special utilities are responsible for purging files whose release dates have passed, and for clearing volumes and reclaiming fragmented space.

Strategies of the NCAR MSS

In most cases data written to the MSS is routed to disk. This data is duplicated onto archive tape within five days. This results in a form of temporary backup until the file migration program deletes the disk copy.

Cartridge volume space allocation and formats are handled by the MSS. All cartridge volumes in the MSS are preformatted with non-MVS volume labels which allow the cartridges to be mounted on nonspecific drives. Each volume has a header block with a copy of the Master File Directory (MFD) entry for that volume. Each file on a volume has a header and trailer block with a copy of the MFD entry for that file (before and after write). The MSS uses standard MVS space allocation for disk volumes. Disk volumes and disk files do not have the specal MSS header blocks.

Each file has a checksum stored as its last block. MSS includes a utility to verify the data on randomly selected archive volumes using this checksum. Historically, this utility has enabled NCAR to predict deteriorating media and to detect unreadable data before a client discovers the problem.

Strategies of CFS

Besides the 3380 disks and the 3480 archive tape systems, CFS has a 3480 active tape system where very large files are written directly from the network to tape cartridges. These active tapes are stored near the tape drives for faster mounting and would be candidates for automation. CFS also supports a variety of other storage systems, including various disk systems, round tapes, the IBM 3850 MSS, the Automated Tape Library for round tape and the MASSTOR M860. Standard MVS access methods are used.

CFS uses bitmaps to handle the space allocation for all volumes. The size of the allocation unit can be varied for different disk volumes to give storage efficiency. A disk file on CFS can be stored in five non-contiguous extents. All tape volumes within CFS use standard MVS labels.

All files within CFS consist of one or more data blocks. Each block is labeled with information that identifies the file to which the data belongs and the classification of the data. The label is verified before the data block is transmitted over the network. The label may also contain a checksum for each block, for error detection and recovery at the data block level.

File Movement Strategies

It is with file movement that CFS and the NCAR MSS differ most significantly. The NCAR MSS allows the supercomputers to directly read and write the storage systems. This approach gives higher data transfer rates and allows a smaller control processor to be used. CFS requires all storage system access to be through the control processor which provides better integrity and security.

File Movement in CFS

The CFS file movement strategy is to simultaneously transmit many files using multiple direct connections to the Los Alamos File Transport System. All supercomputers and some of the other network servers are also directly connected to the File Transport System. Other machines and servers are connected through gateways. The File Transport System is a number of Gould computers connected in a ring to provide a high performance store and forward network. This File Transport System offers the advantages of having large memory to buffer transmissions, and the intelligence to restrict communications and thus enforce security partitioning in the network. Burst rates of 50 Mb/sec are supported.

Los Alamos developed protocols and networking software provide message passing, process-to-process communication, end-to-end file transmission, and remote procedure calls. Implementation of the networking software allows any machine to communicate with any other machine within the security rules, and to use the services of all network servers including CFS. DECNET networks of VAX distributed processors and TCP/IP networks of scientific workstations and personal computers are connected through gateways to the File Transport System.

The CFS software also supports the Network Systems HYPERchannel™ network which is used by most CFS sites either with NETEX™ or with direct drivers. The CDC Loosely Coupled Network and the Cray Superlink are used by one CFS site, and another CFS site uses the MASSTOR MASSNET™ protocol.

File Movement in the NCAR MSS

There are two strategies used to move files within the NCAR MSS. The first services supercomputers via the LDN, the second services network nodes with small file requirements on the MASnet. The LDN allows up to six parallel data streams to be moving data. Each stream can move data at a 3 MB/sec. burst rate for a total system bandwidth of 18 MB.

The file movement through the MSS to a supercomputer starts with a request being presented to the COS MSS control kernel via the COS ACQUIRE/DISPOSE mechanism. This kernel makes certain low-level parameter checks and forms a message packet which is sent to the mass storage control processor via a Network Systems HYPERchannel™ adapter connection. The processor accepts the message, makes further validity checks and queues the request for processing. When resources are available the medium is mounted and positioned, if necessary, and the data path to be used in the LDN is assigned. A message is returned to the supercomputer requesting data be moved via the LDN between a specified device and CPU local disk. When finished, control is returned to the requesting client. If, during a read request, the second access has occurred during a two-day period the data will be migrated from archive storage to the disk farm thereby increasing the chance of the data being read from an online device the third time it is needed.

A file moving through MASnet follows a similar procedure to the LDN path. A request is made from a node on MASnet and received by the MSS. Next, all parameters are verified and a data path prepared. A write to the MSS requires space be allocated on the disk farm. A read from the MSS requires that the data be resident on the disk farm. Once a path has been prepared, data is moved between the disk farm and MASnet. When finished, the MFD is updated and the request is acknowledged complete.

Performance

Both the NCAR MSS and CFS provide a high level of performance, reliability and availability. The MSS uses fast, dedicated paths with the supercomputers while CFS relies on being able to transmit many files in parallel. Both systems have had very good experiences with IBM 3380 disk and 3480 tape. Neither site has lost a file due to 3380 hardware or media problems and only a few files have been lost from 3480 hardware and media problems. Both systems use multiple controllers and multiple paths to provide a high degree of redundancy. Failing equipment is automatically or manually removed from operation without affecting system availability.

NCAR MSS Performance

At the time the MSS was planned and designed, several studies indicated a need for a high bandwidth data path [5]. As compute power, memory, and local disk storage increased, the requirement for high-bandwidth also increased. A relatively small number of large files occupy most of the MSS. Moving these data between a CPU and a device becomes significant. The LDN has provided these transmission rates. We see a maximum data rate of 20 Mb/sec. with an average data rate of 10 Mb/sec. When data is moved over the MASnet the maximum data rate is about 3.5 Mb/sec. with an average rate of less than 1 Mb/sec. The above rates are measured from the time the bitfile mover is granted permission to move data until the request is satisfied.

CFS Performance

The most important CFS performance requirement is to satisfy requests from interactive users in a timely manner. Currently, 97 percent of the 24,000 file transfer requests that typically occur each day are satisfied from disk, where, for machines connected to the File Transport System, the average response is five seconds from the time of the request until the file has been transmitted. The response time depends on the file size and the transmission rate which can vary from 10 Mb/sec for supercomputers to 10 kilobits/sec for personal computers. For the supercomputers, file transmission rates of 5 Mb/sec for disk and 8 Mb/sec for tape are typical.

The most demanding performance requirement is when a large number of graphics files are retrieved from CFS for analysis on the supercomputers. For instance, it is common for a designer to transfer a gigabyte of data which at 5 megabits/sec would take almost 30 minutes. Users improve their response time by sending multiple requests to CFS and transmitting three or four files in parallel.

To keep pace with increasing supercomputer capacity and problem requirements, a goal is to double the CFS file transmission rate in the coming year. The IBM 3990 cache controller will mask most of the disk rotational delay and increase the burst data transfer to 36 Mb/sec. Also, changes will be made to more easily allow users to transmit multiple files simultaneously. The current CFS throughput of about 100 megabits/sec can be increased to provide as much parallelism as is needed.

Future Directions

The future of mass storage systems such as CFS and the NCAR MSS will be to provide storage for extremely large amounts of data and increase the bandwidth at which large amounts of data can be moved. New storage media such as optical disk, optical tape, VHS tape, 8mm tape and any exotic technology yet to be developed will be considered. The media must have capacities on the order of hundreds of gigabytes. Files this large will be generated by future supercomputers. Data transfer rates will have to be over 100 MB/sec in order to move the large files. As file sizes grow, the current top-level error rates of 10^{-14} will have to be improved. Storage system interfaces must be made standard to allow their easy use and interchange by mass storage systems. Current magnetic devices will have to be improved in the near future in order to process the immediate needs of the mass storage community. Robotics for archival storage devices are needed in order to process the large number of mounts across a randomly accessed, large capacity storage.

Future Directions for CFS

The future directions of CFS are being driven by the future directions of the Los Alamos Computing Network. As part of the movement toward the use of standard network protocols and the use of UNIX as a standard operating system, the TCP/IP suite will be implemented in CFS and new CFS interfaces will be developed based on UNIX syntax and conventions. As the use of workstations grows, better CFS interfaces that exploit the graphics capabilities of workstations will be developed, and implementations will be done to provide better storage solutions for an environment of integrated supercomputers and workstations. For instance, a CFS front-end could be developed to provide a Sun Network File System interface to CFS. Ways to provide file transparency throughout the user's environment will be investigated.

The acquisition of more powerful supercomputers, the availability of higher performance graphics and the implementation of a very high-speed network based on the Los Alamos proposed High-Speed Channel (HSC) [6] will greatly increase the CFS file access and file storage requirements. To meet these requirements, high-performance storage systems such as disk arrays and VHS tape systems will have to be acquired. It is likely that these storage systems will be managed by CFS but the data will be accessed directly by the supercomputers. This would have to be done in such a manner that the stringent CFS security and integrity requirements are not compromised. A planned first step is to give supercomputers direct access to the CFS 3480 tape drives.

Future Directions in the NCAR MSS

There are plans to provide an interface for the TCP/IP which would allow the File Transfer Protocol (FTP) access to the NCAR MSS. Plans are also under consideration to provide a process to allow devices compatible with the Small Computer System Interface (SCSI) standards. These devices will not be attached to the MSS LDN. A separate access will be provided by attaching an IBM RT to the IBM 4381 and using its SCSI compatible port. This will allow access to many forms of media including tapes such as 8mm and VHS tapes. In addition to these media, there are slow speed optical disks currently available on the market that are interfaced on SCSI channels. An RT access to SCSI ports promises to provide a flexible and speedy path for clients to introduce data into the MSS.

Future media storage systems will be increasing in capacity and reliability. New disks are already available for replacing the 120 GB disk farm which will increase storage capacity to 240 GBs and increase the bandwidth to 4.5 MB/sec. These devices are available from IBM and other manufacturers and are compatible to current IBM channel specifications.

References

[1] B. Collins and C. Mexal, "The Los Alamos common file system," Tutorial Notes, Ninth IEEE Symposium on Mass Storage Systems, October, 1988.

[2] S. Miller, "IEEE reference model for mass storage systems, version 2.0," SRI Project 4211, Technical Report, 4211-87-TR208, October 1987.

[3] M. Nelson, D. Kitts, J. Merrill, and E. Harano, "The NCAR mass storage system," Digest of Papers, Eighth IEEE Symposium on Mass Storage Systems, pp. 12-20, May 11-14, 1987.

[4] COS Version 1, Reference Manual, SR-0011, Cray Research, Inc., Mendota Heights, Minnesota, 1986.

[5] P. Rotar and R. Jenne, "Mass storage system requirements and projections," Proceedings, Third Annual Computer Users Conference, NCAR Technical Note (NCAR-TN/219+PROC), pp. 90-107, November, 1983.

[6] D. Tolmie, "High speed channel (HSC)," ANSI Draft X3T9.3/88-023, June 29, 1988.