LA-UR -83-2678

CONF-8309165--1

LA-UR--83-2678

DE84 001068

TITLE: A NETWORK FILE-STORAGE SYSTEM

AUTHOR(S) M. W. Collins
Marjorie J. Devaney
Emily W. Willbanks

SUBMITTED TO National Center for Atmospheric Research Scientific Computing Division Third Annual Computer Users Conference
Boulder, Colorado
September 20-21, 1983

DISCLAIMER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

A NETWORK FILE-STORAGE SYSTEM

by

William Collins, Marjorie Devaney, Emily Willbanks
Los Alamos National Laboratory

ABSTRACT

The Central File System (CFS) is a file management and file storage system for the Los Alamos National Laboratory's computer network. The CFS is organized as a hierarchical storage system: active files are stored on fast-access storage devices; larger, less active files are stored on slower, less expensive devices; and archival files are stored offline. Files are automatically moved between the various classes of storage by a file migration program that analyzes file activity, file size and storage device capabilities. This has resulted in a cost-effective system that provides both fast access and large data storage capability.

INTRODUCTION

The Central File System (CFS) is a file storage system for a local network of 40 computers having 5 different operating systems running 24 hours a day, 7 days a week. CFS has been operational since June 1979 in a computing environment that is primarily timesharing, scientific, Fortran, and oriented toward sequential files. Any file stored in CFS can be retrieved on any of the network computers providing the user has access to the file. The CFS storage devices are operated as a storage hierarchy with CFS deciding which storage device each file should reside on. The CFS works very well in providing quick access to a large amount of data. The user interface to the CFS is simple and flexible and was designed for optimum performance with an interactive terminal user. The connection between CFS and the computer operating systems is minimal, a better approach than a previous system, which was integrated with the operating systems.

THE NETWORK ENVIRONMENT

A functional view of the Los Alamos Computing Network is shown in Fig. 1. A locally developed network switch connects the network components. Although most computer networks use the contention bus type of interconnection, such as the Network Systems Hyperchannel, Los Alamos uses the network switch concept because of the intelligence and the buffering it provides. The intelligence is necessary to meet the security requirements and the buffering facilitates the data transmission. A source machine can send data to the network switch without the destination machine needing to be ready to accept the data. In effect, memory in the network switch replaces the more expensive memory buffers in the Network machines.

The principal worker computers in the Network are five Cray-1 supercomputers and four CDC 7600s. Smaller worker computers, such as the DEC VAX 11/780 and the CDC Cyber 825, are also used. The XNET gateway system connects remotely located distributed processors to the Network.

The philosophy of the Los Alamos Network is that the worker computers provide computational services and that special network computers provide support services. A communication system of over 2500 terminals provides interactive access, including remote job submittal, for about 4500 users. PAGES, the print and graphics output system, provides an online capability for print, microfiche, film, and plotted output. FOCUS, the facility for operations control and utilization statistics, provides for the monitoring and control of the worker computers, including the control of production jobs. A planned tape station will provide for off-site data storage and data interchange. The CFS serves as the permanent storage system for most of the network files. The major worker computers do not have permanent file systems. User files are deleted from these machines if they are not accessed within 17 hours. This allows the worker computer disk to be used strictly for active files.

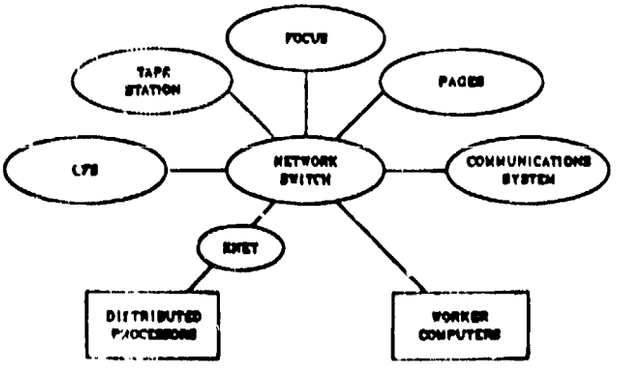


Fig. 1. Los Alamos computing network.

Los Alamos operates under stringent security requirements and it was necessary to partition the network into a Secure network for classified computing, a closed network for administrative computing, and an Open network for computing by any validated user.

Most of the network components, including the worker machines and the terminals, are physically placed in one of the three networks. The CFS, however, is logically partitioned by software to provide file storage for all three networks. The CFS also provides for a controlled file sharing between the three networks.

Los Alamos has had a computer network, including a centralized file storage system, since 1971. The previous file storage system was based on the IBM 1360 photodigital store (a trillion bit storage system).

THE CFS ENVIRONMENT

The CFS configuration is shown in Figure 2. Two IBM 4341 computers serve as primary and backup control processors. All CFS production programs run in the primary control processor. If the primary processor fails, the production programs can be switched to the backup in a matter of minutes. Because the CFS must always be available to the network, a separate means of testing changes is necessary. Therefore, the backup processor is used to run test versions of the CFS programs. The test system programs can be driven from the worker machines or from a network simulator program.

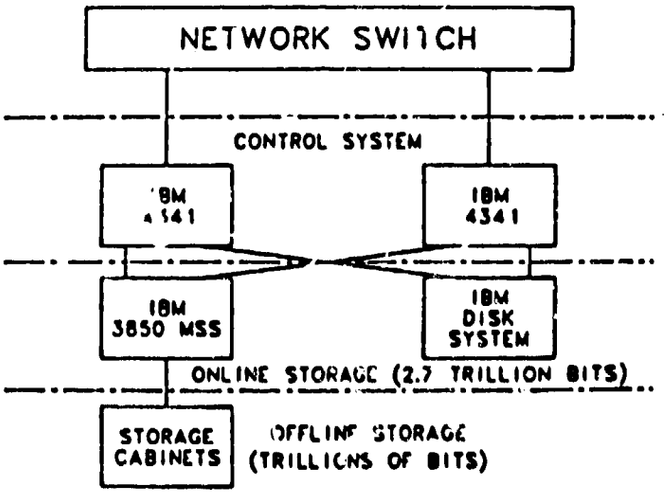


Fig. 2 Central File System configuration.

The CFS programs run as application programs under the IBM MVS operating system. No changes to the operating system were necessary. Most of the CFS program modules are written in PL/1, with Assembler being used for only a few modules. The CFS software has been exported to other installations with similar mass storage devices.

The CFS uses two online storage systems: an IBM disk system and an IBM 3850 Mass Storage System (MSS) with an online capacity of about 2.7 trillion bits. The MSS is a virtual disk system that uses tape cartridges to store data. Under normal operation, the MSS requires no manual intervention because the tape cartridges are automatically moved between the storage cells and the data read/write devices. An offline or archival storage capability is provided by ejecting cartridges containing inactive files from the MSS and storing them in cabinets. The offline storage provides an essentially unlimited amount of space for archival data. Access to offline data does require manual intervention. The tape cartridge is used for archival storage because it has been very reliable and it is easy to reuse space from deleted files since the cartridges are addressable like disk. In 5 years of experience with the tape cartridge there has been no evidence of media life problems.

The disk, MSS, and offline storage are organized as a storage hierarchy. Active files are stored on disk, less active and larger files are stored in the MSS, and inactive files are stored offline on archival cartridges. Files are automatically circulated within the storage hierarchy by a file migration program that analyzes file activity, file size, and storage device capabilities. The average user response time is 5 seconds for disk, 90 seconds for MSS, and 5 minutes for offline storage. The response time includes the access time and the data transmission time. Besides having a higher access time, most files on the MSS are larger than those on disk, so they have a longer transmission time. The average disk file transmission is 300,000 bytes while the average MSS file transmission is 3 million bytes.

The CFS interfaces to the network through a standard file transmission protocol used for actual shipping of files and through a user interface that communicates with the CFS control processors using a set of precisely defined functions. The network has no direct access to any of the CFS storage devices. This restriction is necessary for the Los Alamos security environment, but also has the effect of shielding the network from most CFS changes. In particular, new storage systems can be added and the existing storage systems modified without affecting the network or users.

USER INTERFACE

The CFS user interface is designed to be powerful and convenient for the interactive terminal user while at the same time not requiring a complicated interface with the network computers. The user functions that the CFS provides the network computers follow.

- CREATE a root directory node
- ADD a directory node
- REMOVE a directory node
- SAVE a file
- GET a file

- REPLACE a file
- COPY a file
- DELETE a file
- MODIFY directory information
- LIST directory information
- STATUS of system or user request
- ABORT a request
- MOVE a subtree

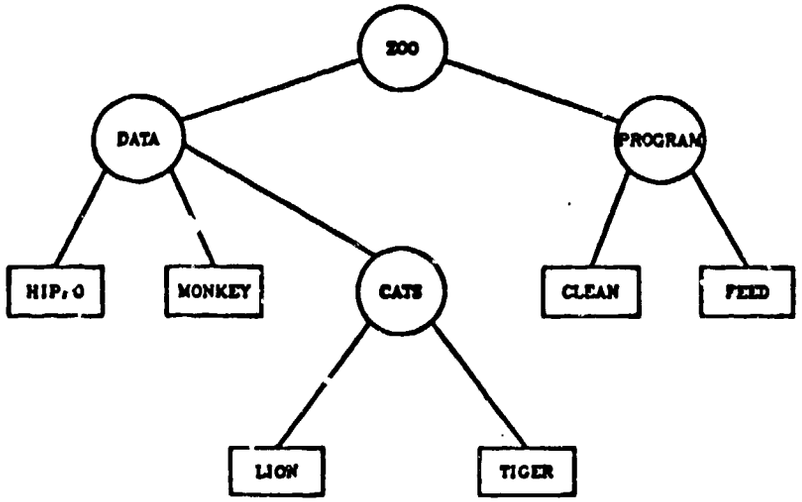
The user interface to CFS is implemented as a standard application level utility called MASS that runs on all the network computers. MASS can easily be put on new computers because it does not require a complicated integration with the operating system. The user, not CFS or the operating system, decides when to store, retrieve, convert, and back up files. Higher level utilities that call MASS are available for repetitive applications that use a fixed sequence of operations. It is much better to put the specialization at the application level than to bury it in the file storage system or the operating system.

Most users' knowledge of the CFS goes no deeper than the MASS utility. The users store and retrieve information as named files and need have no knowledge of CFS storage devices because the CFS determines where the files reside. However, the user can specify that files be placed on separate physical devices. A use frequency can also be specified that influences the initial location of the file. Later, the File Migration Program detects the actual activity and places the file on the most suitable device to maximize user convenience and to minimize user expense, regardless of any user specified values. This device independence not only makes it convenient for users but also enables CFS to make the best use of its resources and to obtain the best possible performance.

No direct access or manipulation of CFS data is allowed. When the user wishes to access data in a file, the complete file is transmitted from the CFS to the worker machine disk where it becomes just another worker machine file. If a user wants to replace a file that has been changed, the complete file must be transmitted back to the CFS.

The CFS maintains a tree structured directory that allows users to organize their data in a logical and reasonable manner. Users can organize their own directory trees in a manner consistent with their needs and abilities; they can use the tree structure but are not required to do so. Figure 3 is an example of a simple tree structure. The circles are directory nodes and the boxes are file descriptor nodes. The nodes exist as keyed records in the CFS Master Directory. These nodes contain a substantial amount of information. Directory nodes contain user access information and a list of descendant nodes (which can be either directory or file descriptor nodes). File descriptor nodes contain the physical location of the file, user access information, and file activity

information. Reference to a node is by its unique path name. For example, the path name of the file LION is ZOO/DATA/CATS/LION. Each node has its path name as the record key so the nodes can be retrieved without traversing the tree. A user can create as many tree structures as is desired.



ZOO/DATA/CATS/LION

Fig. 3. Tree structure example.

Access to nodes is strictly controlled, but users can easily be given different access privileges to different parts of a tree. For example, a user could be given read access to all files in the tree by an appropriate entry in the directory ZOO and also be given write access to just the file LION by an appropriate entry in the file descriptor node for LION.

The tree-structured directory also offers the opportunity for users to perform operations on groups of files and/or directories. For example, GET all the files in a subtree, DELETE all the files in a subtree, LIST information for a subtree, or MODIFY information for a subtree.

No restrictions are placed on the amount of storage users can have, but users are charged for file accesses, the amount of data transferred, and the amount of data they have stored. Access charges are higher and storage charges are lower for ofiline files.

FILE MIGRATION

File migration is based on the premise that it is possible to improve user response, optimize storage use, and minimize user costs by automatically moving files to storage devices that fit the characteristics of their actual usage history and current size. The CFS migrates files between the disk, online MSS, and offline storage.

File migration is performed by a CFS application program that scans the Master Directory to decide which type of storage device each file should reside on. This File Migration Program then requests the File Management Program to move the necessary files. A typical run takes 4 to 12 hours (depending on how much data must be moved and how busy the File Management Program is) and moves 2500 files or 8 billion bits of data. The File Migration Program reads the complete CFS Master Directory and calculates a priority for each file. The priority of a file, as shown in Figure 4, is a function of its activity (aged access count) and size. A new aged access count is calculated by multiplying the old aged access count by the aging factor of 0.9 raised to a power that is the number of days since the last calculation. Whenever a file access occurs, a new aged access count is calculated and then incremented by one.

$$P = \text{FUNCTION}\left(\frac{A}{S}\right)$$

P IS FILE PRIORITY
A IS AGED ACCESS COUNT
S IS FILE SIZE

$$A_{NEW} = A_{OLD}\,(F)^{P}$$

F IS AGING FACTOR = 0.9
P IS PERIODS SINCE LAST CALCULATION

Fig. 4. File priority calculations.

Migration between disk and MSS is based on a priority value selected so that if all files having a priority greater than this value are stored on disk, the disk will be filled to 93% of capacity. This leaves 7% of the total disk space for new files. Files greater than 64 million bits currently are not stored on disk.

The online to offline migration of files is based on idle time (days since last reference), file size, and the user-specified activity. The idle time limits are automatically adjusted to move data off line to match the rate at which users are accumulating data, keeping the online MSS at 94% capacity. Currently if the user specifies the file to be online and the file size is small, it will not be migrated to an archival volume until the idle time is greater than 360 days. For the maximum file size, the idle time has to be only 100 days. Between these two limits the idle time is a logarithmic function of the file size. If the user changes the file use frequency from online to archival, lower idle times of 45 days for small files and 15 days for large files are used.

Offline files are migrated online when their access count is greater than a fixed value. Depending on their priority, the offline files may be moved to either disk or the MSS.

Once migrated, a file will not be migrated again until a specified time passed. The migration program also purges expired files and can move all files from a specified device

to reduce fragmentation (not a major problem) or to allow device maintenance.

PERFORMANCE

In August 1983, after 4 years of operation, users have stored one million files, which occupy 12 trillion bits of storage. The growth is currently 380,000 files and 4.6 trillion bits per year. If the storage space were not reusable, the growth would be over 10 trillion bits per year. The typical daily activity is 13,000 file accesses with 90 billion bits of data transferred, 1500 file deletions, 3500 lists, and 200 modifications. The prime shift usage averages 1200 file accesses with 7 billion bits of data transferred per hour. The CFS availability to users is better than 99%.

The CFS storage hierarchy and file migration process provide quick access to a large amount of data. Table I shows the excellent CFS response.

Table I. Storage Hierarchy Performance

Device	File Type	% of Storage	% of Accesses	Typical Response
Disk	Active	<1	87	5 s
3850 MSS	Less Active Larger	16	12	90 s
Offline Storage	Inactive	83	1	5 min

Disk storage is used for less than 1% of the data stored, but about 87% of the 13,000 daily file accesses are to disk. More than 80% of the data is stored offline but only about 1% of the accesses are to offline files.

CONCLUSIONS

The CFS is a valuable and critical resource of the Los Alamos computer network. Its storage hierarchy allows fast access to a large amount of online storage plus a cost-effective archival storage. This is done using mass storage technologies that are available, proven, and reliable. Stringent security and integrity requirements are met while allowing flexible data sharing in a multicomputer network. The analysis of file statistics and the migration of files improves performance and minimizes costs for the users. Additional benefits include:

- Economy of scale. More storage can be provided at less cost with a centralized system.

- Files ar available to all computers in the network without maintaining multiple copies.

- The single CFS Master Directory provides good management and control of data. Usage, accounting, and descriptive information are available for every file.

- A very reliable file system.

- A device-independent interface for the users. Storage devices can be added or changed without affecting the network.

- A large reduction in the use of conventional magnetic tape. Tape mounts are only one-tenth of what they were in 1976, and the tape library size has been reduced 75% despite a substantial increase in computing power, activity, and the number of users.