LA-UR -81-3558

LA-UR--81-3558

DE82 006108

TITLE: A NETWORK FILE-STORAGE SYSTEM

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

MASTER

SUBMITTED TO Spring Compcon '82 Digest

DISCLAIMER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

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
Los Alamos, NM

Abstract

The Common File System (CFS) is a file management and mass 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 (over five trillion bits currently stored).

Introduction

The Common File System (CFS) is a centralized file storage system for a local network of 28 computers having 5 different operating systems running 24 hours a day, 7 days a week. CFS has been operational for 2 1/2 years in a computing environment that is primarily timesharing, scientific, Fortran, and oriented toward sequential files. The CFS serves as the permanent storage system for most of the network data as the major network computers do not have permanent file systems. 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

The CFS serves as the Los Alamos Computer network file storage system. A functional view of this network is shown in Figure 1. The network has a core of worker machines that currently includes four Cray-1S supercomputers, three CDC 7600s, and one CDC 6600 with a combined capability of over 200 million floating point operations per second (equivalent to eighty IBM 370/168s). Smaller machines such as the DEC VAX 11-780 and the CDC Cyber 73s are used as special purpose worker computers. All of the worker machines run timesharing operating systems. They are front-ended by a communications network of over 1500 terminals and are back-ended by a file transport network that includes the CFS file storage system, the Print and Graphics Express Station (PAGES) output system, and the Extended Network Access System (XNET) that connects many distributed processors to the network. The file network also provides for worker-to-worker file and message shipping. The CFS serves as the permanent storage system for most of the network data. The Cray and 7600 machines do not have permanent file systems. Their user files are deleted if they are not accessed within 17 hours and not even system files are saved across cold deadstarts. The high performance disk on these machines essentially serves as a disk cache.

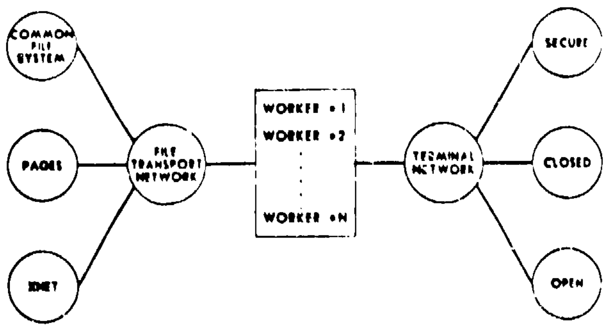

Fig. 1. Los Alamos computer network.

Los Alamos operates under very 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 where any valid user can compute. 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, for 10 years. The previous file storage system was based on the IBM 1360 photo digital store that was 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 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.

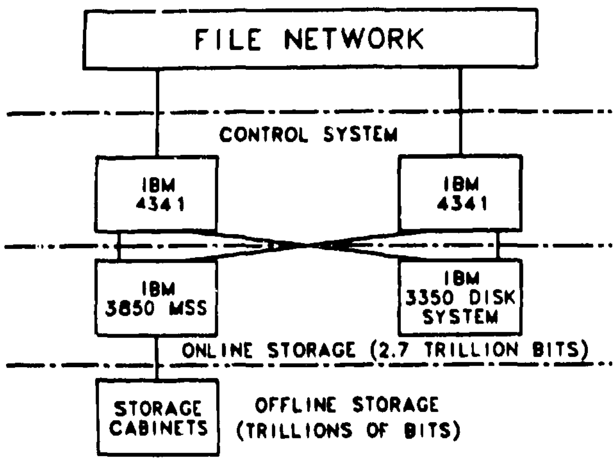

Fig. 2. Common file system configuration.

The CFS programs run as application programs with no modifications to the IBM operating system. Almost all of these 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 systems.

The CFS utilizes two online storage systems: an IBM 3350 disk system with a capacity of 50 billion bits and an IBM 3850 Mass Storage System (MSS) with an online capacity of about 2.7 trillion bits. The MSS uses tape cartridges to store data. Under normal operation, the MSS requires no manual intervention as 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 3350 disk, MSS, and offline storage are organized as a storage hierarchy. Active files are stored on 3350 disk, less active and larger files are stored in the MSS, and inactive files are stored offline on archival cartridges.

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 impacting 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.

- o CREATE a root directory node
- o ADD a directory node
- o REMOVE a directory node
- o SAVE a file
- o GET a file
- o REPLACE a file
- o COPY a file
- o DELETE a file
- o MODIFY directory information
- o LIST directory information
- o STATUS of system or user request
- o ABORT a request

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 MASS or CFS, 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, as 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 (online or offline). 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.

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.

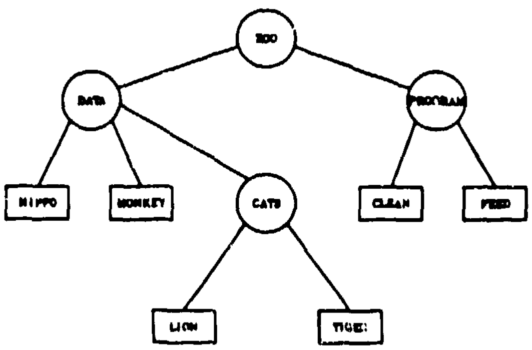

Fig. 3. Tree structure example.

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 offline 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 uses a storage hierarchy of 3350 disk, online MSS, and offline storage and migrates files between these devices.

File migration is performed by a CFS application program that cycles through the steps of deciding 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 cycle takes 4 to 12 hours depending on how much data must be moved and how busy the File Management Program is. The File Migration Program reads the complete CFS Master Directory and calculates a priority for each file, where the priority increases with activity and decreases with size. 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 which is the number of days since the last access. Whenever a file access occurs, a new aged access count is calculated and then incremented by one.

P = function (A/S)
P is file priority
A is aged access count
S is file size

$$A_{new} = A_{old}\,(F)^{D}$$

F is the aging factor = 0.9
D is days since last access

Fig. 4. File priority calculation.

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 90% of capacity. This leaves 10% of the total disk space for new files. In a typical week the CFS will migrate approximately 7000 files and 20 billion bits of data between disk and MSS. Files greater than 60 million bits are not migrated to disk.

The online to offline migration of files is based on idle time (days since last reference), file size, and the user specified activity. If the user specified 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 270 days. For the maximum file size, the idle time only has to be 180 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 or if the file is specified to be online for a few weeks and then archival, lower idle times of 45 days for small files and 15 days for large files are used. The idle time limits are automatically adjusted to move data offline to match the rate at which users are accumulating data, currently about 50 billion bits of data each week.

Offline files are migrated online when their priority 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 week has 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 early December 1981, after 2 1/2 years of operation, users had stored 480,000 files, which occupy 5.1 trillion bits of storage. The growth is currently 235,000 files and 2.5 trillion bits per year. The typical daily activity is 12,000 file accesses with 60 billion bits of data transferred, 1000 file deletions, 3000 lists, and 200 modifications. The peak period usage is 1200 file accesses with 5 billion bits of data transferred per hour.

The CFS storage hierarchy and file migration process provides quick access to a large amount of data. Table 1 shows the excellent CFS response. Disk storage is used for less than 1% of the 5.1 trillion bits of data stored, but about 85% of the 12,000 daily file accesses are to disk. More than three-fourths of the data is stored offline but less than 1% of the accesses are to offline files. To maintain the CFS performance with increased file access and file storage requirements resulting from additional computing capacity, it is planned to install a large IBM 3380 disk system in 1982. This disk system will allow more files to be stored on disk and will prevent the MSS from becoming saturated.

Device	File Type	% of Storage	% of Accesses	Typical Response
3350 Disk	Active	1	85	5 sec
3850 MSS	Less Active Larger	19	14	1 min
Offline Storage	Inactive	80	1	5 min

TABLE I STORAGE HIERARCHY PERFORMANCE

In 1981, CFS was available to the users 99% of the time. The scheduled down time (0.4%) had little impact on the network as it was done during off hours when there were few terminal users and production jobs could retrieve their CFS files and be started before the CFS shutdown. The unscheduled downtime, which could affect the network performance, was 0.6%. So far no files are known to have been lost or destroyed because of CFS software problems, and only about 50 files have been destroyed by various hardware problems, primarily problems with MSS cartridges.

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 is improving performance and minimizing costs for the users. The following benefits have resulted from this approach.

- Economy of scale. More storage can be provided at less cost with a centralized system.
- Files are available to all computers in the network without maintaining multiple copies.
- The single CFS Master Directory provides for good management and control of data. Usage, accounting, and descriptive information are available for every file.
- A very reliable file system. The only file losses have been caused by hardware problems.
- A device-independent interface for the users. Storage devices can be added or changed without affecting the network.
- The necessary file security for the Los Alamos environment.
- A large reduction in the use of conventional magnetic tape. Tape mounts are only one-tenth of what they were five years ago, and the tape library size has been cut to one-third despite a substantial increase in computing power, activity, and the number of users.