N93-16323

EMASS™: AN EXPANDABLE SOLUTION FOR NASA SPACE DATA STORAGE NEEDS

Anthony L. Peterson
P. Larry Cardwell

E-Systems, Inc. Garland Division
Dallas, Texas

Abstract

The data acquisition, distribution, processing and archiving requirements of NASA and other U. S. Government data centers present significant data management challenges that must be met in the 1990's. The Earth Observing System (EOS) project alone is expected to generate daily data volumes greater than 2 Terabytes (2×10^{12} Bytes). As the scientific community makes use of this data their work product will result in larger, increasingly complex data sets to be further exploited and managed. The challenge for data storage systems is to satisfy the initial data management requirements with cost effective solutions that provide for planned growth. This paper describes the expandable architecture of the E-Systems Modular Automated Storage System (EMASS™), a mass storage system which is designed to support NASA's data capture, storage, distribution and management requirements into the 21st century.

Introduction

We first discuss NASA's requirements for mass storage with a focus on functional and performance specifications. Next, an overview of the EMASS architecture is presented and evaluated with respect to NASA's requirements. The major EMASS architectural components, hardware, software and interfaces, are then explored with emphasis on the data management capabilities of the EMASS software.

NASA Requirements

Requirements for large volume, mass storage systems have been well established in order to meet the storage needs for NASA's space and Earth science information systems. The use of sophisticated data acquisition instrumentation will continue to evolve, providing large, increasingly complex data sets to be processed, distributed and archived. Therefore, data storage requirements

will continue to grow nonlinearly through the 1990's. For example, the Earth Observing System (EOS) project alone, generating daily volumes greater than 2 Terabytes, will require automated storage libraries with capacities greater than 500 Terabytes by the late 1990's. E-Systems is also currently developing storage systems to meet existing U. S. Government and commercial requirements to be delivered in 1993 having automated data storage library capacities greater than 200 Terabytes.

Data management requirements such as these within NASA and other U. S. Government data centers present significant challenges that must be met in the development of new mass storage systems. These systems must meet increasing performance requirements with cost effective solutions while providing for planned growth. E-Systems is developing the EMASS architecture to address these requirements for extremely large, expandable data storage and data management systems.

As we view NASA's supercomputer-based data management systems we see a need for high bandwidth, high density tape recorder systems having the data quality characteristics of a computer peripheral. As scientific data processing requirements move towards open systems environments, the file management software and server should support a UNIX environment. The file management software structure should provide application specific integrated data management solutions. A file server with high I/O bandwidth is required to accommodate simultaneous data transfers from multiple high bandwidth tape recorder systems. Finally, to keep a perspective on hardware and maintenance costs, the use of commercially available equipment is strongly emphasized.

EMASS Architecture Overview

The EMASS architecture is a family of hardware and software modules which are selected and combined to meet these data storage requirements. Figure 1 illustrates the EMASS architecture. EMASS is a UNIX-based hierarchical file management system utilizing both magnetic disk and tape. The storage capacity ranges from one to several thousand Terabytes depending on the type of storage library used. It has the capability to support both a graphical and metadata interface to the user. The system is user driven by standard UNIX and unique EMASS commands and has user configurable automatic file migration. The EMASS system employs standard protocols for user file transfer, communications and network interfaces.

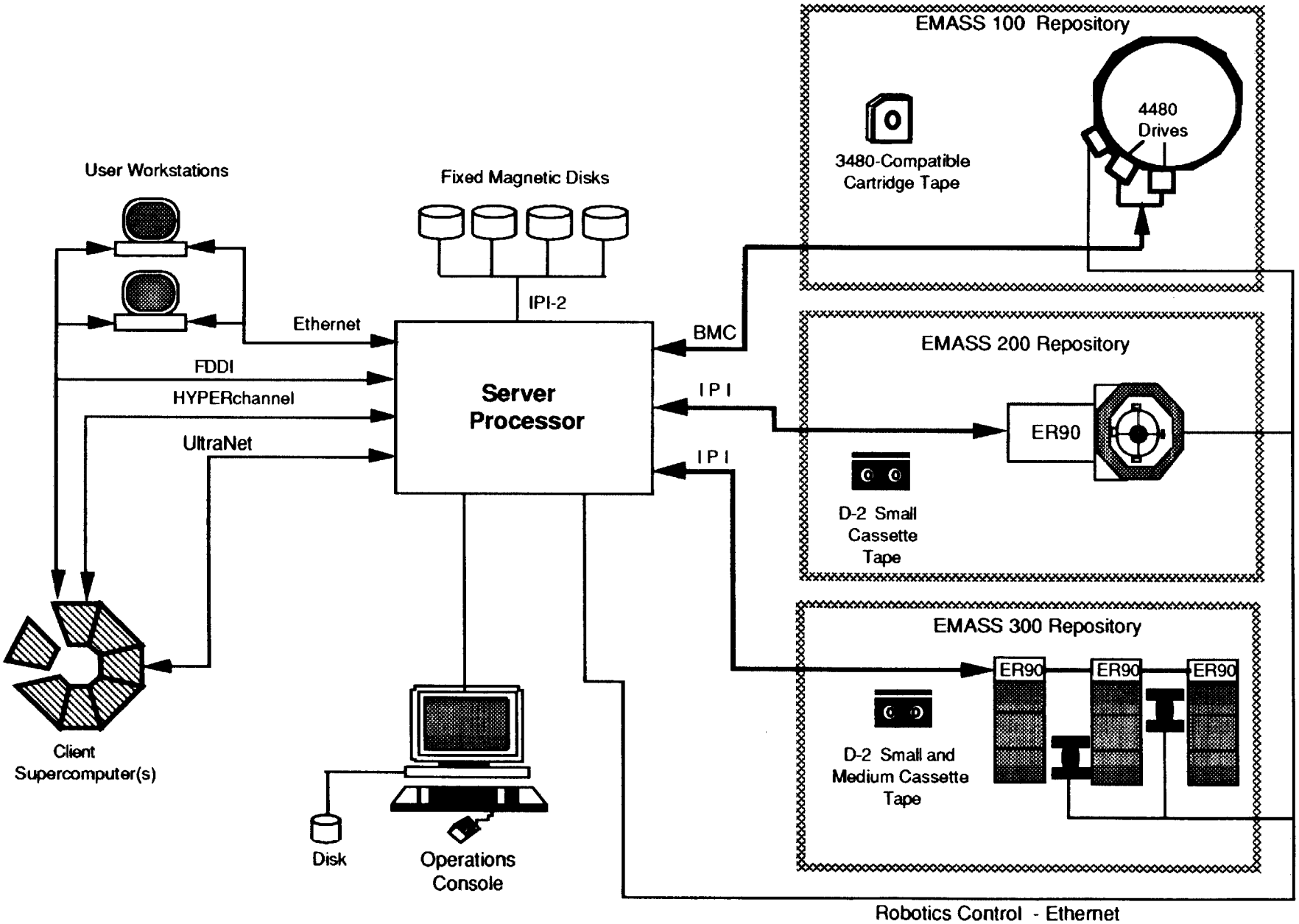

Figure 1. EMASS Architecture

The EMASS system operates as a large data storage node on a network, servicing client requests over a number of standard interfaces, including Ethernet, FDDI, DECnet, HYPERchannel™ and UltraNet™. The system is a two-level hierarchical data storage system. Magnetic disk is the first level of storage, and magnetic tape is the second. Data is managed via a selectable migration policy based on data class, a method of data segregation addressed in a subsequent section. Two alternative types of magnetic tape for data storage are included in the EMASS design: 3480 tape cartridges and D2 digital tape cassettes. Files are migrated to 3480 or D2 tape depending on the migration policy for the data class to which that file belongs. The physical volume repository (Miller[1]) functionality is implemented in three separate types of storage libraries which are selected based on user requirements for performance and expandability.

EMASS Hardware

The storage system file server function is implemented in a CONVEX C3200 series computer. The CONVEX was selected after an extensive survey of available computers. The major evaluation factor leading to the selection of the CONVEX machine is its high I/O throughput performance. The CONVEX supports four channels, each having I/O bandwidths of 80 Megabytes per second peak and 60 Megabytes per second average. Other key evaluation factors included cost, compatibility with a UNIX environment, modularity, expandability, upgradability, reliability, and support.

The file server interfaces with three types of tape libraries, the STORAGETEK (STK) 4400 Automated Cartridge System, the EMASS DataTower™ and the EMASS DataLibrary™. The STK tape cartridge library data interface is implemented through ANSI standard Block Multiplexor Channel interfaces which connect to the STK 4480 drives through a SUN Library Server . The DataTower™ and DataLibrary™ data interfaces are implemented with enhanced ANSI standard IPI-3 tape controllers within the file server connected to E-Systems ER90 digital D2 recorders.

The DataTower and DataLibrary robotic systems provide data archive expandability. The DataTower, with dimensions illustrated in Figure 2, serves as a medium scale storage device, with a capacity of 6 Terabytes on 227 small D2 cassette tapes. This device was implemented by

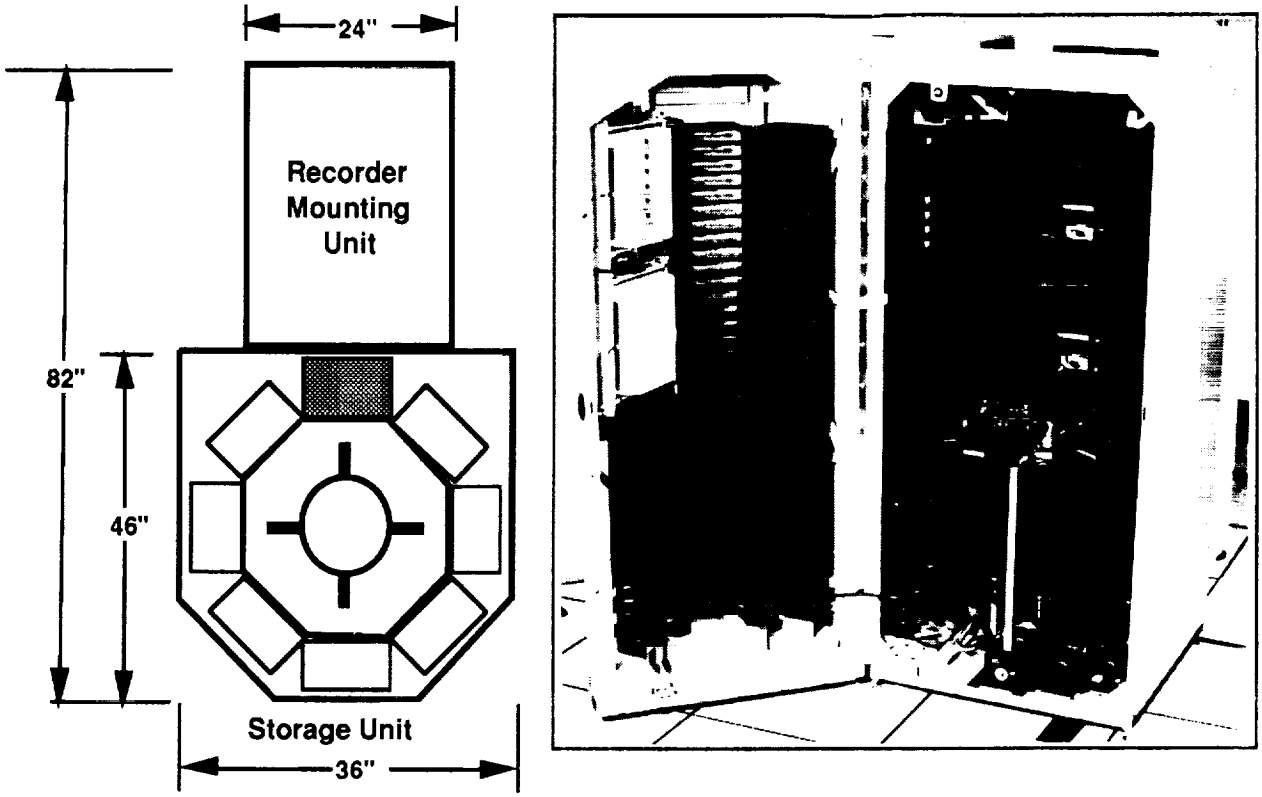

Figure 2. EMASS DataTower

modifying an existing automated robotics tower currently in volume production for the broadcast industry. The device may be expanded by adding up to three additional expansion storage units, for a total capacity of 25 Terabytes.

The DataLibrary, illustrated in Figure 3, is a modular aisle architecture comprised of a series of modules each four feet in length. This design specifically addresses the needs for a modular, expandable data storage solution required for NASA's large data archives from EOS and Space Station Freedom. Each shelf module contains up to 207 small, or 192 medium, D2 tape cassettes, for a maximum capacity of 14 Terabytes. Shelf units reside in rows on either side of a self-propelled robot and can be added incrementally as the library grows. The row of modules may be expanded to lengths of 80 feet, providing a maximum of 288 Terabytes per row. Further expansion is accomplished by adding additional rows and robots. Cassette access times are specified at 45 seconds maximum for robot travel spanning an 80 foot aisle for cassette retrieval. The DataLibrary configuration will be housed within a sealed watertight structure with interior fire protection using CO_2 supplied on demand.

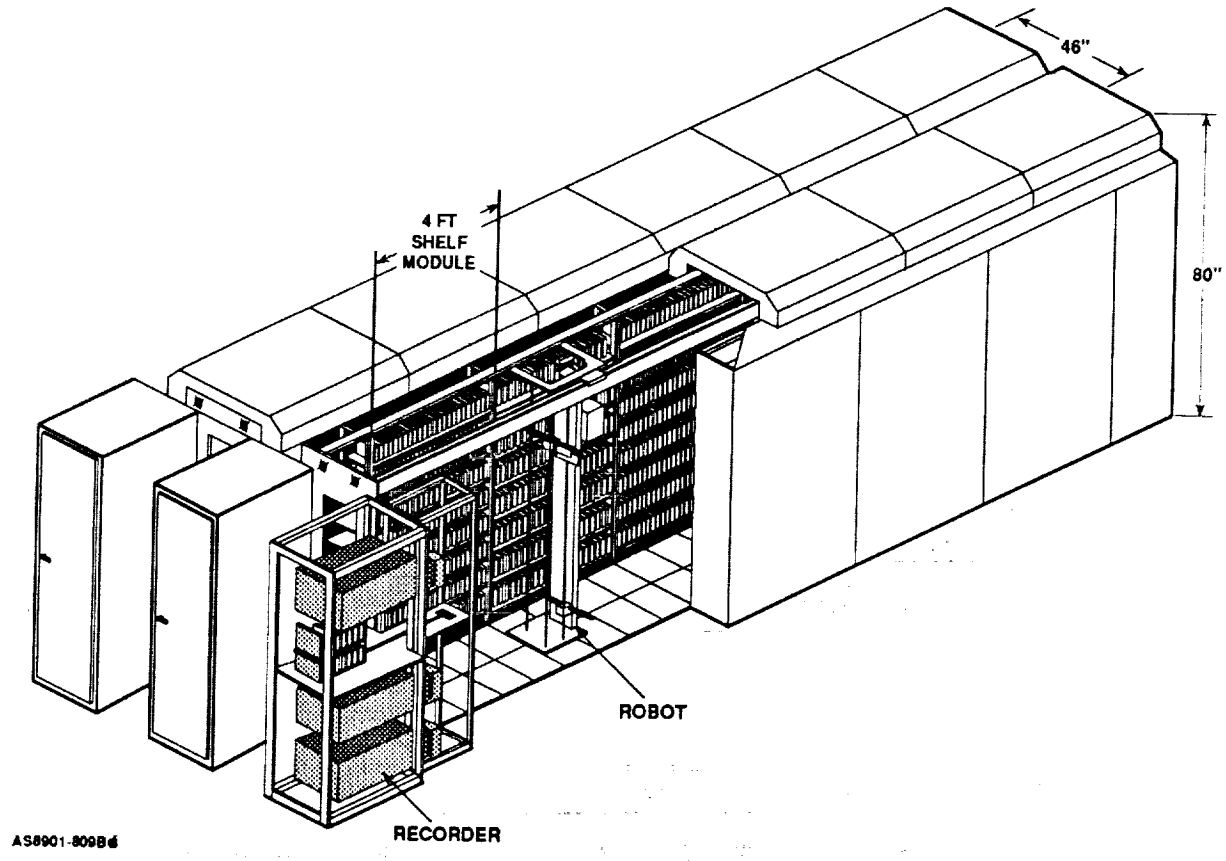

Figure 3. EMASS DataLibrary

E-Systems selected D2 helical scan tape and recorder system technology to meet high density and bandwidth requirements. As shown in Table 1, the 19mm D2 tape cassette is available in three form factors: small, with a capacity of 25 Gigabytes of user data, medium, with a capacity of 75 Gigabytes and large, with a capacity of 165 Gigabytes.

The suitability of the 19mm D2 helical scan media and recorder for use as a computer peripheral has been reviewed by Wood[2]. The D2 recorder provides a format already in wide use within the broadcast industry. Ampex, SONY, and Hitachi have delivered over 2000 D2 units to the broadcast industry since 1988. The D2 video broadcast recorder has been modified to develop the ER90 digital recorder peripheral. Key features of the ER90 recorder include air guides to minimize tape wear, azimuth recording and automatic scan tracking.

The ER90 provides a sustained data rate of 15 Megabytes per second, with burst rates up to 20 Megabytes per second. Additional error detection and correction coding has been implemented using a three-level interleaved Reed-Solomon code. Resulting error event rates of 1 in 10^{13} bits are being achieved. The recorder absolute positioning velocity is 300 inches per second and logical positioning velocity is 150 inches per second. This results in an average file access time for mounted media of 10 seconds for D2 small cassettes and 30 seconds for D2 medium cassettes. To provide compatibility with existing file management systems the ER90 provides ANSI 9-Track file labeling compatibility.

PARAMETER	SPECIFICATION
Tape Media	19mm - D2
Tape Cassette Capacities	25 GB (S), 75 GB (M), 165 GB (L)
Data Rate	15 MB/sec - Sustained 20 MB/sec - Burst
Error Detection/Correction	3-Level Interleaved R-S Code
Error Event Rate	1 in 10^{13} bits
Tape Positioning Velocity	300 in/sec - Absolute Address
Average File Access Time (Mounted Media)	10 sec - D2 Small 30 sec - D2 Medium
Data Format	Compatible With ANSI 9-Track File Labeling
Peripheral Interface	Enhanced IPI Physical (ANSI X3T9/88-82) IPI-3 Logical (ANSI X3.147-198?)

Table 1. Recorder System Performance

The ER90 drive uses the enhanced IPI physical interface (ANSI X3T9/88-82) and the IPI-3 Magnetic Tape Command Set (ANSI X3.147-1988)at the logical interface level. The enhanced IPI physical interface can sustain transfer rates commensurate with the basic transport performance. A second enhanced IPI interface port can be added to allow a separate master-slave path to another server. A large internal buffer (approximately 60 Megabytes) has been incorporated for rate smoothing to minimize recorder start-stop sequences.

EMASS Software

The EMASS server stores files in an extended UNIX File System (UFS). EMASS software is divided into separate components as depicted in Figure 4. These components are the user interface, the event daemon, the migration manager, the file mover, and the physical device manager. All EMASS software executes as UNIX processes at the application level. All UNIX kernel enhancements/modifications were accomplished by CONVEX and are included in ConvexOS™ 9.0.

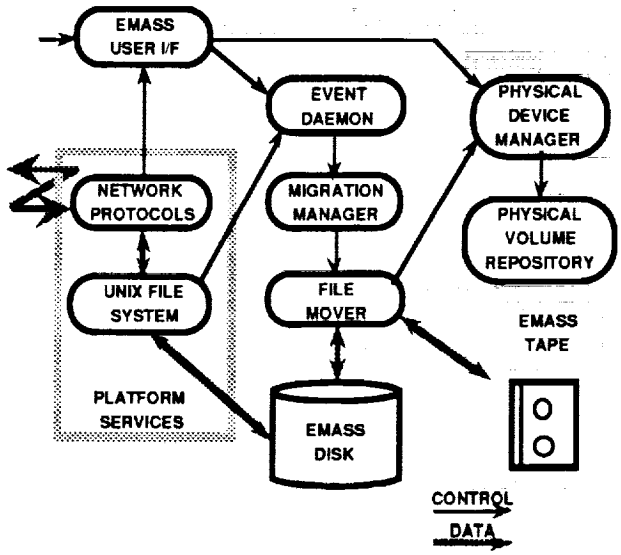

Figure 4. Overview of EMASS Software.

Users have two methods to gain access of EMASS migration services. One method is through a user interface front-end which provides migration override control to end-users. The other method is through direct access of the CONVEX UFS. This second method provides transparent access to EMASS file migration services to both local and networked users.

In order to provide the ability to transparently migrate files, CONVEX has upgraded their ConvexOS to allow the UFS to provide notification of selected critical file system events to a user-level event daemon. This modification is similar to those made by the BRL/USNA Migration Project (BUMP)[3], a joint development of the US Army Ballistic Research Laboratory and the US Naval Academy. The EMASS event daemon receives file events and forwards them to the migration manager. The migration manager collects this information. When migration policy is triggered, the migration manager will select files for migration and forward the list of selected files to the file mover.

The movement of files from an EMASS server disk to magnetic tape and from magnetic tape to an EMASS server disk is controlled by the file mover. For each list of files to migrate, a file mover process is created to perform the read and write operations. The file mover design provides for the addition of software routines to support new media types.

The final major EMASS software component is the physical device manager. The physical device manager provides a standard interface for tape movement services to all other EMASS software components. The physical device manager will translate a generic tape movement request into the format required by the target physical volume repository (PVR). The translated request is then sent to the PVR for processing. The physical device manager will later receive the results from the commanded PVR. The results are then placed into a generic format and sent to the process that requested the tape movement. Additional PVRs will be supported by the addition of software modules to the physical device manager.

DataClass™

Before describing EMASS software in more detail, a discussion of the abstraction known as DataClass™ is required. The file systems that are to be provided EMASS migration services are subdivided at specific points in the file system tree structure by identifying those directory point(s) beneath which all files are to be managed alike. These directory point(s), which are referred to as migration directories, are what define each DataClass. When a new directory point is added to a DataClass, the event daemon will request the UFS to associate the directory and all files below it (both present and future) with the event daemon.

Figure 5 depicts a DataClass to migration directory relationship. The directory /test/dick/special is the only migration directory in DataClass SPECIAL. All files beneath /test/dick/special will be managed together. The DataClass PURPLE contains all files under the directories /prod/blue and /prod/red, but none under /prod/green, showing that some directories at a certain level may be excluded from a DataClass. All files under /test/jane and /test/dick/public belong to DataClass TESTERS. This illustrates that the assignment of migration directories to DataClass is not restricted to a certain level in the tree structure. In fact, migration directories from different file systems may be in the same DataClass. Also, a file system can be mounted onto a mount point underneath a migration directory, for example /test/jane/dir1.

The definition of DataClass is key to site administration. Migration policy parameters are configurable on a DataClass basis, thus providing the EMASS administrator with a great deal of control over the behavior of the EMASS system. Time interval between policy application, time required on disk prior to migration, and desired time for migrated files to remain cached on disk are examples of DataClass based migration policy parameters. Quotas for tape utilization (both

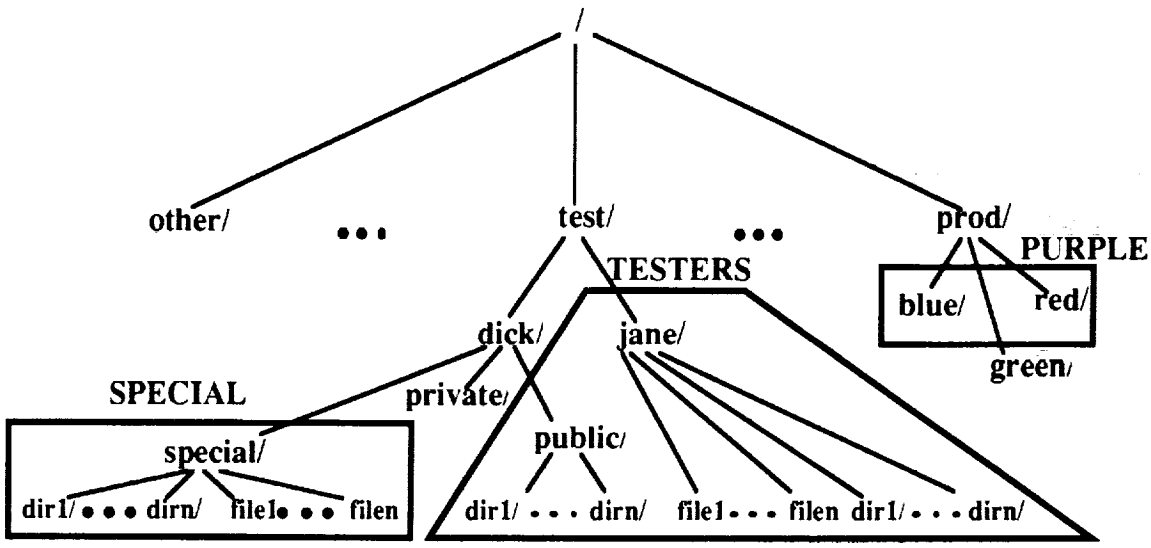

Figure 5. DataClass Example.

a warning limit and a hard limit) are also kept on a DataClass basis. DataClass based parameters are kept in the INGRES database, so tuning can be easily done while the EMASS system is active.

EMASS software also uses DataClass as the means to segregate files on tape. All files on a tape will be of the same DataClass. This provides a level of physical security for those sites which might require it. This segregation also ensures that retrieval of files from different user groups (as defined by DataClass) will not collide trying to access the same physical tape.

EMASS Interfaces

A key concept of the EMASS architecture is that it will provide multiple archival file storage choices to the users of various networked client systems. EMASS supports connectivity over industry standard interfaces including Ethernet, DECNet, FDDI, HYPERchannel, and UltraNet. This provides connectivity from the smallest workstations to the largest supercomputers. The industry standard transfer protocols available to the user include the File Transfer Protocol (FTP)

from the TCP/IP family, the Network File System (NFS™) as defined by Sun Microsystems, and the UNIX utilities UNIX to UNIX copy (UUCP) and remote copy (rcp). This support is provided by placing the EMASS interface under the UFS. The EMASS system will receive notification of all managed file system events for files in every DataClass. Migration services are therefore provided for any connectivity available on the EMASS server system to the UFS. Thus, as new connectivity options are offered by the EMASS server vendor, the EMASS system will automatically also provide support.

Hierarchical Data Storage

EMASS provides three levels of data storage. These three levels of hierarchical storage are EMASS server disk, robotically managed tape, and human managed tape. When files are placed in a DataClass, the residency is on server disk. The EMASS migration policy will schedule placing the files onto tape based on the migration rules defined for its DataClass. The user can preempt the migration policy by giving the EMASS system a directive to migrate specific files to tape immediately.

When the migration policy is executed for a DataClass, all files in that DataClass which are not solely on tape are examined. If the time since last modification (or the time since retrieval from tape if unmodified) of a file is greater than that specified in the policy for that DataClass, then that file's data is placed in the staging directory. If an up-to-date copy of the file is not on tape, the file is added to a list of files to be migrated. When all files in the DataClass have been examined, the list of files to migrate is forwarded to the file mover.

To store files on tape, the file mover first allocates tape(s) through the physical device manager. The files in the list are next migrated to tape. The file mover will record the location of the tape copy of each file in the INGRES database as they are successfully written to tape. The disk copy is left intact as a cached copy which will later be removed from disk when either 1) the length of time since migration has exceeded its DataClass defined limit or 2) the file system requires additional disk data blocks and it is the oldest staged file on disk. When the file mover has completed migrating files to tape, the tape and drive are released back to the physical device manager.

Not removing disk data blocks from the staging directory on strictly a first-in-first-out basis provides additional flexibility. A file system can be divided into DataClasses which can have different access requirements. Each DataClass will have its own disk retention period, so one

portion of a file system can have data that is not cached on disk as long as another portion. One DataClass thus can be set up to never free disk data blocks except when required to provide needed free disk space for the file system.

When the UFS receives a request for data blocks for a file which is not currently on disk, the requesting process is suspended and the EMASS event daemon is notified. The event daemon forwards that notification to the migration manager which immediately instructs the file mover to retrieve the requested file to disk. The file mover requests the tape containing the file be mounted and copies the file to disk. The requesting process is now allowed to continue processing. The EMASS system maintains knowledge of the tape copy. If the disk copy is unchanged, the file will not be re-migrated by the migration policy.

The retrieval of a user-directed range of bytes from a file is also available to the EMASS user. This is accomplished much like the retrieval of a complete file. However, the file mover will copy the specified range of bytes into a UNIX file of a different name as specified by the user. Thus, the user can retrieve only the portion of the file of interest, reducing the amount of data brought back.

The EMASS system will also manage tapes that are not under robotic control. This will allow sites to have EMASS management of many more tapes than the robotic system can support. When access is requested for a file that is only on a tape that is not under robotic control, the EMASS system will request the operator to return the tape to active service so that the file may be copied onto disk. The effect to the user is only a longer delay waiting for a tape mount.

Infinite File Life

A mass storage system must provide for the integrity of its client's data. In order to insure that the client's data is always available, the EMASS system has several features to provide safeguards against data loss. These features include automatic Error Detection and Correction (EDAC) monitoring and secondary file copy maintenance.

For every file segment written by an ER90 drive, the drive will automatically perform a read while write comparison. If the data written is not recoverable or to successfully recover the data written required more than a minimum threshold of correction, that segment is automatically re-written to tape by the drive without any action required by the host system. This provides positive assurance that the data written is retrievable without much stress on the EDAC at the time it is recorded.

For every file read by an ER90 drive, the EMASS system will request the drive to return EDAC statistics and if the level of correction was excessive, that tape will be placed in a "suspect" list for system administrator action. The system administrator can then at a later time request the EMASS system to move all files off of the old tape. This provides for refreshing the EDAC encoding for all files that were on the old tape.

To ensure the health of D2 tapes that have not been accessed for a while, the EMASS system provides a tape sniffing service. Tape sniffing is the process of periodic monitoring of tapes that have not been accessed for a length of time defined by the data center. The EMASS system will schedule the reading of sample files from the tape and then examine the EDAC statistics to determine if the level of correction was excessive. If excessive, the tape is placed in the "suspect" list for system administrator action.

As an added measure of protection for tape-based files, secondary file copying is provided. If enabled for its DataClass, files will automatically have a secondary copy maintained on a separate tape. This DataClass feature can be overridden on a file basis, thus allowing the user to request a secondary copy be created when the DataClass default is to not maintain a copy. The user can conversely request the EMASS system not to maintain a secondary copy of a file when its DataClass default is to maintain a copy.

Through the use of automatic EDAC monitoring and secondary copies, the EMASS system provides for the integrity of its client's data. The life of the EMASS client's data can in fact be prolonged well beyond that of any one type of storage media, as the file sniffing service will promote data from one media onto another.

Summary

EMASS software provides a UNIX-based data storage solution with automatic and transparent file migration and retrieval. Data archive centers can be provided with very large (up to Petabytes), expandable, automated data storage systems. These data storage systems connect to high speed networks, providing 24 hour per day accessibility for rapid delivery of requested data in the Space Station and EOS era. File access can be provided to networked users through standard file transfer protocols. A graphical user interface can also be provided. Thus, the client is not required to have special networking software. The implementation of DataClass provides a flexible method for tuning the behavior of the system at each installed center.

Trademarks

EMASS, DataTower, DataLibrary and DataClass are trademarks of E-Systems, Inc.

CONVEX and ConvexOS are trademarks of CONVEX Computer Corporation.

UNIX is a trademark of AT&T.

INGRES is a trademark of ASK Computer Systems.

NFS is a trademark of Sun Microsystems, Inc.

HYPERchannel is a trademark of Network Systems Corporation.

UltraNet is a trademark of Ultra Network Technologies, Inc.

References

[1] S.W. Miller, "A Reference Model for Mass Storage Systems", *Advances in Computers,* Vol. 27, Academic Press, 1988, pp. 157-210.

[2] Tracy G. Wood, "A Survey of DCRSi and D2 Technology", *Digest of Papers,* Proc. Tenth IEEE Symposium on Mass Storage Systems, May 1990, p. 46 (1990).

[3] Michael John Muuss, Terry Slattery, and Donald F. Merritt, "BUMP, the BRL/USNA Migration Project", *Unix and Supercomputers*, 1988.