Data Migration Facility Development Update

Thomas W. Lanzatella and *Neil Bannister,* Cray Research, Inc

ABSTRACT: *This paper illustrates current work in Data Migration Facility. We describe requirements for a hierarchical storage management solution suitable for a Cray Research machine environment. Release history and current status are covered followed by a description of future plans. Over the next two years, DMF will be infused with new database technology and its architecture will change to accommodate distributed capabilities without requiring the UNICOS Shared File System as an interconnect mechanism.*

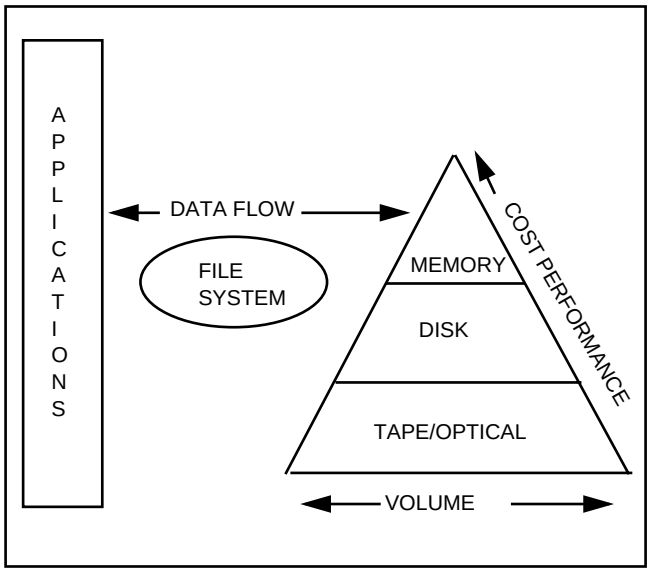
Introduction

The primary objective of a hierarchical storage management system (HSM) is to preserve the economic value of storage media and stored data. The high I/O bandwidth of a Cray Research machine environment is sufficient to overrun online disk resources. Consequently, capacity scheduling in the form of native file system migration has evolved into an integral part of many customer environments and is a requirement for the effective use of Cray computational tools. Besides ensuring that adequate disk space is always available, capacity scheduling affords the ability to maintain a data space that is larger than the online disk resource. Oversubscription leads to a requirement recognizing the value of stored data as the same or higher than that of data which is online.

The HSM provided on Cray systems must support a range of storage management applications. In some environments, HSM is used purely to manage highly stressed online disk resources. In other environments, HSM is an organizational tool for safely managing large volumes of offline data. In all environments, the Cray HSM must scale to the storage application and to the characteristics of the available storage devices.

In performing its function, the Cray HSM must transport large volumes of data on behalf of many users. As system interrupts and occasional storage device failures are all but unavoidable, the Cray HSM must provide the ability to validate the storage environment. Positive, verifiable safety of data is a requirement for Cray HSM.

The Cray Data Migration Facility (DMF) has evolved around requirements for scalability and safety of data. As a file system migrator, DMF provides the ability to manage the capacity of online disk resources by transparently moving file data from disk to offline media. Most commonly, offline media is tape managed by the UNICOS Tape Subsystem but can be any bulk-storage device accessible by FTP. Transparent migration means that a user cannot determine, using POSIX-compliant commands for file system enquiry, that a file is online or offline. Only by using special command options can the actual residence of a file be determined. In performing its function, DMF leaves inodes and directories intact within the native file system.

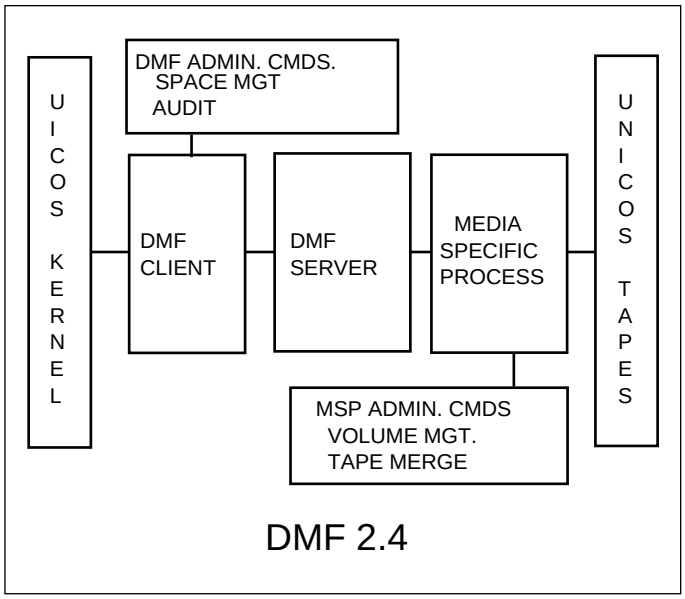


The DMF administrator determines how disk capacity is managed by selecting which file systems are managed by DMF and the volume of free space that will be maintained on each file system. DMF space management begins with a list of user files that are ranked by administrator-defined criteria. File size and file age are the most commonly employed ranking criteria. File migration occurs in two stages. Files are first migrated to offline media. Once an offline copy is secure the file is eligible to have its data blocks released. The administrator chooses both the percent of file system volume to migrate and the volume of free

space. File migration can be triggered by administrator as part of automatic space management or by the file owner using manual migration requests.

Offline media is the destination of all migrated data. The part of DMF that manages offline media is called the Media Specific Process (MSP). The MSP manages a pool of media volumes and is the agent that moves file system data to and from offline media in response to migration requests. The MSP is designed to make full use of high-capacity, compressible media and to handle a huge volume of transactions. The data recording format employs blocking and checksumming to ensure accuracy of data and to facilitate recovery in the event of media failure.

DMF provides several capabilities that enhance the safety of operations and the integrity offline data. The DMF administrator can configure multiple instances of the MSP with each managing its own pool of media volumes. DMF can be configured so that file system data is migrated to multiple, offline locations. Much of the work done by DMF involves transaction processing that is recorded in databases. Database transactions are journaled so that in the event of an unscheduled interrupt, the journal files can be used to rectify database inconsistencies arising from incomplete transactions. Besides journaling, a tool is provided that is used to audit for inconsistencies between the file system and the DMF databases.



Cray DMF Architecture

DMF consists of a client, a server and a Media Specific Process. The DMF-Client accepts requests from the DMF administrator or from users to migrate file system data and communicates with the UNICOS kernel to maintain the migration state of a file in the file inode. The DMF-Server accepts requests from the DMF-Client. The server is responsible for dispensing a unique identifier for each file that is migrated. The server also determines the destination of migration data and forms requests to the appropriate MSP(s) to make offline copies of the data. The MSP accepts requests from the server. For out-bound data, the MSP accrues requests until a sufficient volume of data is available to justify a volume mount. Requests for data retrieval are satisfied as they arrive. When multiple retrieval requests involve the same volume, all file data is retrieved in a single pass across the volume.

DMF Release History

Over the last two years, DMF has evolved in response to new data recording hardware, new configurations for Cray clusters linked by UNICOS Shared File System and to elevated requirements for reliability. DMF 2.2 was released in November, 1994. The development theme for DMF 2.2 was a reimplementation of the tape MSP. The result of this work is an MSP that handles compressible tape media and absolute block positioning. Other benefits include full use of asynchronous I/O for both tape and disk transfers.

DMF 2.3 was developed to support clusters of Cray machines linked by the UNICOS Shared File System. In DMF 2.3, the architecture of DMF was changed to a client/server model. The purpose of this division is to provide migration services on a Cray machine that is different from the machine where the DMF server is running and where bulk storage resources are attached. Furthermore, DMF 2.3 supports the use of multiple instances of the DMF server, each maintaining its own set of MSPs and databases. Most typically, a DMF server is configured on each machine where bulk storage resources are connected. Reliability and failsafe operations are an important part of tightly-coupled Cray clusters. DMF 2.3 is designed to function in a failsafe manner. It is possible to specify a sequence of machines, in priority order, which act as backup platforms for the DMF server in the event of a machine or network failure. Another advantage of separating DMF into client and server is that because the client runs independent of the server, the server can be taken down for regular maintenance while leaving the client up. In this manner, user processes making requests for migrated data experience normal responses where previously they would have received errors. DMF 2.3 was released during July, 1995.

The theme for DMF 2.4 development is to increase DMF reliability by simplifying the path taken by file data as it moves from the file system to offline media. Previous to DMF 2.4, file data was moved to a special directory managed by DMF until the MSP could be activated to make an offline copy. This directory is called the premigration directory. While this mechanism is effective, it requires special operational consideration when using dump/restore. Inadvertent duplication of premigration directories during file system maintenance can place unmigrated data at risk if documented procedures are not followed precisely.

In DMF 2.4, the UNICOS kernel interface with the DMF client is reorganized to not require the use of premigration files as a temporarily repository for data blocks of migration files. Part of the kernel interface reorganization includes the addition of a new file system called /inode. The UNICOS changes necessary to support DMF 2.4 are part of UNICOS 9.0.2. DMF 2.4 is

self-adapting in that it can run on any version of UNICOS from 8.0 onward. When the UNICOS version is 9.0.2 or above and the /inode file system is mounted, DMF 2.4 does not use premigration files.

In reorganizing the kernel interface to DMF 2.4, a major improvement in automatic space management has also been implemented. Previously, space management was accomplished with three utilities: the DMF controller, dmmctl; the file ranking utility, dmhit and the space freeing utility, dmfree. Because these utilities were all separate processes, it was prone to inaccuracy in that the file ranking order could be different from the order in which space was made available by dmfree. For some sites, migration policy can be deleteriously affected by this inaccuracy. For this reason, dmhit and dmfree have been combined with dmmctl into a single utility. In DMF 2.4, the file ranking order is identical to the order in which files are removed. The internal hitlist generator used by dmmctl executes ten times faster than that used by the dmhit utility.

DMF has grown to be a useful backup service for UNICOS file systems. This was facilitated by the UNICOS versions of the dump and restore utilities which were enhanced to detect that a file had been migrated. By selecting the appropriate dump option, the administrator can cause only unmigrated files to be dumped. If migration policy is coordinated with file system backups by maintaining a 100% migration threshold and by dumping only unmigrated data, file system backup can be pared to a short operation in which only inodes and directories are recorded on the dump media. The attraction for this process is that machine availability is dramatically increased by not having to shutdown to perform file system dumps.

Previous to DMF 2.4, establishing a 100% migration policy conflicted with policies designed to ensure space availability through selection of high-ranking migration candidates. In DMF 2.4 a new utility, dmmigall, has been introduced that allows for implementing a 100% migration policy without infringing on space management policies.

Database journaling is important technology within DMF. In DMF 2.2, as part of the new MSP, database journaling was implemented so as to allow direct application of journal files to an MSP database without first converting the database to source. In DMF 2.4, this same capability has been extended to the DMF server database. A new utility, dmdrecover, has been introduced for this purpose. Using dmdrecover, journal application to the server database is four times faster than under DMF 2.3.

Database auditing is a DMF process that validates the consistency of the server database with respect to the UNICOS file system. The utility that performs database consistency checks is dmaudit. In DMF 2.4, dmaudit runs four times faster than under DMF 2.3.

Since the introduction of support for Absolute Block Positioning (ABP) for tape positioning operations, problems have been experienced with the reliability of this feature. Problems are apparent in Storage Technology controller microcode and, on occasion, in parts of UNICOS. In some cases, these problems have resulted in loss of tape resident data. Field Notices 2014 and 2122 have been issued advising that ABP be disabled until corrections can be supplied for the STK controller microcode and for the J90 (and Y/MP-EL) IOS. In consideration of these problems, the DMF 2.4 tape MSP has been altered to perform a positioning check after each ABP operation where the following operation is a write. For this reason, we advise all DMF customers to upgrade to DMF 2.4 as soon as is practical.

Future Plans for Improved Database Integrity

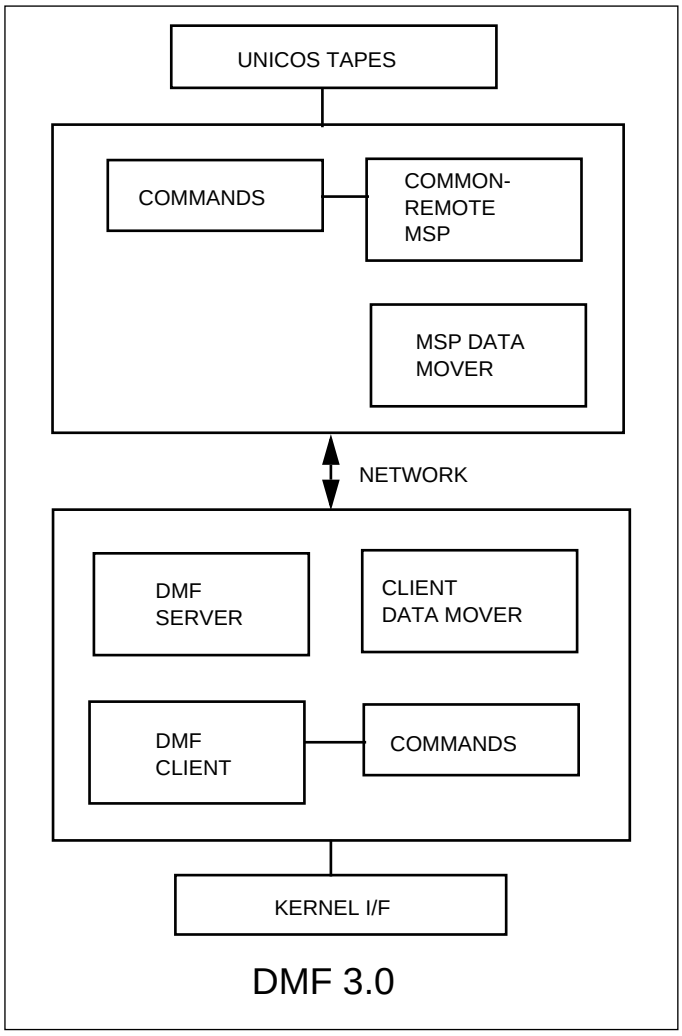
Today, DMF is installed on approximately 200 Cray machines world-wide. More than 25% of all J90s and nearly all high-end machines are placed with DMF. DMF product stability is very high. Software Problem Reports (SPRs) filed against DMF are at the lowest point in its 6-year history. DMF 2.4 is in production use at the Cray Research facility in Eagan where approximately .9 million files are under DMF management. DMF 2.4 represents the conclusion of a 2-year development program to remedy design issues dating from the initial implementation of DMF and to adapt DMF to support prevalent customer configurations.

As the number of DMF installations increases, database technology becomes a determining factor in overall product stability and in the cost of supporting customer environments. While DMF reliability is high, events causing inconsistencies in DMF databases such as unscheduled system interrupts are an increasing burden to support staff. Since DMF has become a critical service at customer sites, time spent repairing databases is time that the Cray machine is not available for production use. We estimate that hundreds of hours of CPU time are spent each year, cumulatively, across the DMF customer base repairing or compressing and rebuilding DMF databases.

For this reason, the development theme for DMF 2.5 will be the implementation of a new database package in DMF. The new package will continue to provide database transaction journaling. Journaling is a means of ensuring exact correspondence between the database and the file system inodes. Beyond journaling, the new database package will employ 2-phase commit technology to ensure that only complete transactions are applied to the database. The new database will also eliminate the need to compress and rebuild the database. The compress and rebuild cycle is a consequence of a problem with the current database package being unable to reallocate holes in the database files that arise from file removal. Pilot studies with the new package show that the performance will be equal or better than the current package.

DMF 2.5 will alter the external interfaces of DMF utilities that perform database operations. Additionally, since the new databases will have a different format than under DMF 2.4 (and before) a conversion process will be required in order to upgrade. A utility will be provided for conversion.

DMF 2.5 will be available during 3Q96. We will drop support for the old tape MSP and for the Station MSP in DMF 2.5.

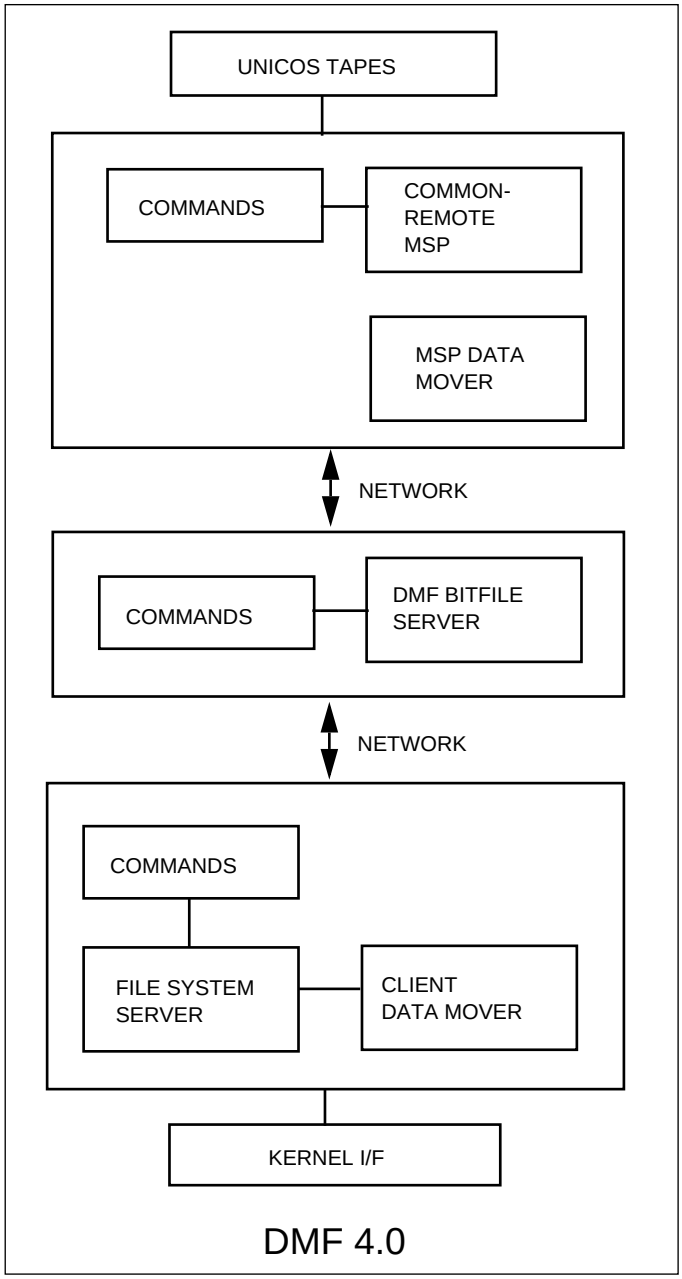


Future Plans for Distributed Services

Distributed DMF supports file system migration in a tightly coupled Cray cluster environment. In this context, distributed migration services are limited to those machines that are linked by UNICOS Shared File System. Current DMF architecture requires that the MSP have direct, native access to a file system under DMF control. In a Cray cluster, an unshared file system on a machine without tape equipment cannot be managed by DMF.

Customer environments are evolving to include multiple Cray machines. In many cases, these machines are geographically dispersed and tape resources are not always present on every machine. While NFS, DFS and FTP are effective ways of sharing data in a distributed environment, they do not address the issue of capacity scheduling for remote file systems.

The development theme for DMF 3.0 is the support of remote UNICOS file system migration. Shared File System will not be a functional requirement. DMF 3.0 will have the ability to migrate file system data from a Cray machine without tape resources, across a network, to a Cray machine where tapes are located. The DMF 3.0 development program will shift DMF architecture to be more consistent with the IEEE Reference Model for Storage Systems. As such, new components and terminology changes will be implemented.



New components for DMF 3.0 are data movers. One mover, the Client Data Mover, will mediate data transfer on behalf of the remote machine where the managed file system is located. Another mover, the MSP Data Mover, will mediated data transfer on the machine where offline data is stored. The mover protocol will employ authentication to insure secure communications, checksumming for data integrity and caching to deal with the asynchronous nature of large file transfers and the foibles of mountable media.

With the introduction of data movers, the tape MSP will change to be functionally detached from the DMF-Server so that

it can execute (optionally) on a remote Cray machine. Additionally, the DMF 3.0 tape MSP will store migrated data on behalf of multiple DMF-Servers. Today, each instance of a DMF-Server requires its own set of MSPs. While useful, this configuration presents operational problems with tape transport scheduling and administrative overhead related to tape merging and volume pool management. The common-remote tape MSP will eliminate these problems. Multiple MSPs continue to be supported and a DMF server will be able to migrate data to more than one (local or remote) MSP.

No command changes are anticipated for DMF 3.0. New configuration options will be supported allowing for remote execution of the tape MSP. No operational changes are anticipated in the manner of DMF administration except those required to initiate and monitor the common-remote tape MSP. We are planning to begin DMF 3.0 field testing during the first-half of 1997.

The development theme for DMF beyond 3.0, tentatively called DMF 4.0, will be to continue enhancing distributed processing capabilities for DMF. In this phase and using the secure infrastructure implemented for DMF 3.0, the DMF-Client will be functionally separated so it can execute remotely from the DMF-Server. The DMF-Client will be renamed to DMF File System Server. The DMF-Server will be renamed to the DMF Bitfile Server.

The advantage of this architecture is that all UNICOS dependencies are isolated to the File System Server where file inode state changes are mediated during the migration process. The Bitfile Server will dispense file identifiers (Bitfile IDs) and determine the destination of migrated data. Under DMF 4.0, client migration platforms will share a common Bitfile Server and therefore the space of Bitfile IDs. Under DMF 4.0, it will be possible to use dump and restore to move user files from machine to machine and have migrated data follow the inode automatically. Under DMF 3.0 this would not be possible because the Bitfile ID space is not shared.

Several other changes are introduced with DMF 4.0. A new dmaudit utility will be provided that can be executed in either the client machine environment or in the Bitfile Server environment. The DMF databases, numbering three today, will be merged into a single database. Many DMF administrative commands will be affected by the distributed nature of DMF 4.0. We are planning to begin DMF 4.0 field testing during the first half of 1998.

Summary

DMF is a leading storage management solution. The product has evolved around requirements for scalability and safety of data. Today, approximately 200 customers are using DMF to address the issues of capacity management and long-term storage of data. Cray customer environments are evolving to include multiple and geographically dispersed machines. In response to new requirements, a 2-year development program has been initiated. The DMF 3.0 product introduces remote file system migration between Cray machines with a common-remote tape MSP. DMF 4.0 introduces separation of the DMF-Client (File System Server) from the DMF-Server (Bitfile Server) and a common Bitfile ID space. The program is expected to conclude in the second half of 1998 with a midpoint delivery of DMF 3.0 during the second half of 1997.