Integrating D3 Technology into an Existing DMF Configuration

Alan K. Powers, Sterling Software, Numerical Aerodynamic Simulation Facility, NASA Ames, Moffett Field, CA

ABSTRACT: *The Numerical Aerodynamics Simulation (NAS) Facility's CRAY C916/1024 has 260+ gigabytes (GB) of Superhome storage with 2 terabytes (TB) of migrated storage. NAS was a beta test site for the Storage Technology Corporation (STK) Redwood device and Cray Research Inc's (CRI) DMF 2.3. This discussion includes a list of resolved problems and various metrics relative to transfer, search, rewind, unload, and mount timings. A discussion of how DMF 2.3, with the advanced tape media-specific process (MSP), will integrate the Redwood tape drives along with the existing STK 4490 tape drives to potentially reduce the number of tapes used by 90 percent.*

NAS C90 Configuration

The NAS Facility currently administers a CRAY C916/1024 (C90) that accesses a "virtual" online file system. The C90 contains 16 CPUs and 1024 megawords (MW) of main memory, 1024 MW of solid-state storage device (SSD), and five I/O clusters. Three communication channel adapters connect the Operator Work Station (OWS) and two FDDI rings. Six HiPPI channel adapters connect four Maximum Strategy RAID[6] devices (Gen IV) and two HIPPI "switched networks". Four tape channel adapters connect four cross-coupled StorageTek (STK) 4490 control units with a total of 16 STK 4490 tape transports inside two Library Storage Modules (LSM, *aka* silo); two more tape channel adapters connect an IBM control unit and eight IBM 3490E manual tape drives.

The C90 uses UNICOS (8.0.3.2) and DMF (2.3.1) to manage 263 GB of online disk and 2 TB of offline storage. The total disk capacity is 338 GB for various support and test file systems and run-time file systems (/big, a 29 GB file system using 1.42 GB of SSD ldcache, and /fast, a 5.2 GB SSD file system).

Superhomes

The 263 GB of migrated storage is shared among five file systems totaling one million files. The support staff file system has 17 GB DD42 disks. The other four 60 GB file systems are for NAS customers and are described as *Superhomes.* Superhomes were created to reduce the number of file systems for the customers, and to increase storage and transfer rates for the home directories. The previous home directories and migrated storage file systems were combined to form the Superhome file systems.

Superhomes are configured with primary and secondary partitions. The primary partition is 8 DD60s used for small (< 1 MB) files and inodes (and files overflowing the secondary partition). The secondary partition is a Gen IV RAID device used for large files (>= 1 MB). Transfer rates of 50+ MB per second to the Gen IV are common. Only files greater than 1 MB are migrated to tape. With this 1 MB limit, more than 85 percent of the files are left online; which greatly reduces the number of records DMF manages and reduces the tape file "hit rate" to less than 5 percent.

CRAY ESCON & Redwood Configuration

For the beta test evaluation, a single ESCON channel adapter (FCA-2), Fiber Optic Extender (FOE-1), new version of ***stknet*** and micro-code (version 1.1) were installed on the CRAY. STK's SD-3 helical tape cartridge subsystem codename Redwood was connected the ESCON channel. The Redwood with two tape drives was silo attached, with one drive dedicated to the CRAY. The Redwood cabinet also includes the controller transport unit for the ESCON fiber connection, unlike the STK 4490 which has a separate cabinet for the control unit. The Redwood's micro-code level used during the test was 1.2 and 1.3. The Redwood's maximum block size is 256 kilobytes (KB) with a tape device buffer size of 64 megabytes (MB).

Redwood Media Description

The D3 tape is the same form factor as the original 3480 tapes. Looking at the tape from the front there are some visible differences:

1. The leader block location is on the opposite (front right) side of the 3480.
2. The write protect location is on the front center instead of the front left of the 3480. The write protect has changed from a thumb wheel to a slide bar.
3. There is a small rectangular notch on the front left; the 3480 does not have a notch.

The label area is also different: there is an additional character space used to detect media capacity and type. Currently, four letters are used: A - 10 GB; B - 25 GB; C - 50 GB; D - Cleaning tape.

Required Silo Software: STKNET and ACSLS

For the tape daemon to use the silo attached Redwood drives a new version of CRI's ***stknet*** and STK's ACSLS (version 5.0) must be installed. The ***stknet*** software is not backward compatible with the previous version of ACSLS, nor is ACSLS compatible with the previous version of ***stknet***.

The ***stknet*** was bundled with the operating system, but now ***stknet*** software is separately licensed and priced. If a site had two CRAYs sharing an ACSLS workstation for the STK 4490 drives and only one CRAY system using the Redwood drives, both systems would need upgraded software and both systems would need licenses.

Some of the differences in ACSLS 5.0 are:

- Mixed transport support for STK 4480, 4490, 9490 and Redwood drives.
- Media type is recognized when the tape volume is entered into the silo.
- The database has switched from using Ingres to Oracle.

Hardware and Software Testing Problems

Problems are expected in any beta test. To help resolve the NAS problems, a weekly teleconference was held to review the previous week's activity. STK was very cooperative and responsive when problems occurred.

Robot Problem

After the Redwood was attached to the silo, the robot hand had problems putting tapes in the silo cell. When the robot placed the tape in the home cell, it was off by about a 1/2 inch. With each try, the robot arm adjusted then tried again until the tape went into the cell. Several actions were undertaken before the problem was resolved: new camera light and a new camera was installed, and the robot arm was adjusted.

Media Detection Problem

ACSLS V5.0 detects the type of media being entered into the silo. Initially, this sometimes failed, depending on which cap door was used. The Redwood tapes were "recognized" as 3490E tapes. After several attempts to resolve this, it was determined that the door was misaligned by 1/8 inch.

Tape Load Problem

When the Redwood was configured in standalone mode (not silo attached), simple tests were successful. After the Redwood drive was silo attached, every time a tape was mounted, the drive sensed *beginning of tape* (BOT) before the tape had finished loading. The silo had very little activity, and the camera would not move after mounting the tape into the drive. The bright lights from the camera caused the Redwood to sense BOT before the physical BOT. This problem was fixed with a Field Change Order (FCO) by an STK hardware engineer.

Buffer Underrun

A test program was written to stream data quickly to the Redwood. The Redwood generated a 'chk 47E5' error (Buffer Underrun) and aborted the transfer. The testing could only continue after resetting the ESCON and Redwood device. This problem was fixed in micro-code level 1.3, but it seemed more like a buffer overflow problem.

Internal Leader Header

An Internal Leader Header (ILH) is part of the Redwood tape format. Besides keeping track of several usage factors, such as the number of read and write passes and mounts, the ILH also has the index search information for logical block id to provide fast search (up to 60 times the read or write speed). When the ILH is not written correctly, the search times become non-linear, as seen in Figure 1(tape T00200). According to STK, this sometimes happens with ASCII label tapes on UNICOS systems. This was still an outstanding problem at the end of the beta test.

Forward Space Block Time-out

After DMF locates a block id, it then does a *Forward Space Block* (FSB) until it reaches the data it needs from tape. To emulate this, a small test program was created to write a full tape and then measured how long it took to scan one 10,000 block unit (2.4 GB). Each time the program tried to scan the first 10,000 blocks, the Redwood tape device would abort, which caused the ESCON channel to fail. When the tape daemon was restarted while the ESCON had this problem, the tape daemon failed. The Redwood and ESCON devices had to be taken out of the tape configuration to restart the tape daemon and the system had to be rebooted to clear the ESCON channel. STK was developing micro-code (1.4) to resolve this problem, which was still an outstanding problem at the end of the beta test.

Redwood Testing

Most of the testing effort was spent on *locate block id*, because this consumes the largest percentage of time retrieving the data. Every site's experience will differ due to the way tests are written and measured.

The Redwood silo mount times are about the same as the 3490E tapes. The ***tpmls*** reported the average mount time to be about 45 seconds for Redwood tape mounts; dismount time was about 20 seconds.

With the advanced tape MSP in DMF 2.2, the tape format has changed to take advantage of *locate block id* (fast seek), compression, and recognizing end of tape. The tape format divides the tape into logical zones. The zone size is a configurable parameter, and the zone is packed with several files called chunks. If a file is written across tapes, it has multiple chunks. For files larger than a zone, the zone size expands to fit the file. The block id is recorded for each zone size.[1]

Several test programs were written to emulate the way in which DMF uses tapes with the new format. The programs tested read, write, *skip block*, and *locate block id*. All the test data read and written to the Redwood tapes used 256 KB blocks. It is helpful to understand the new tape format and know the timings of the tape functions when setting the zone size in order

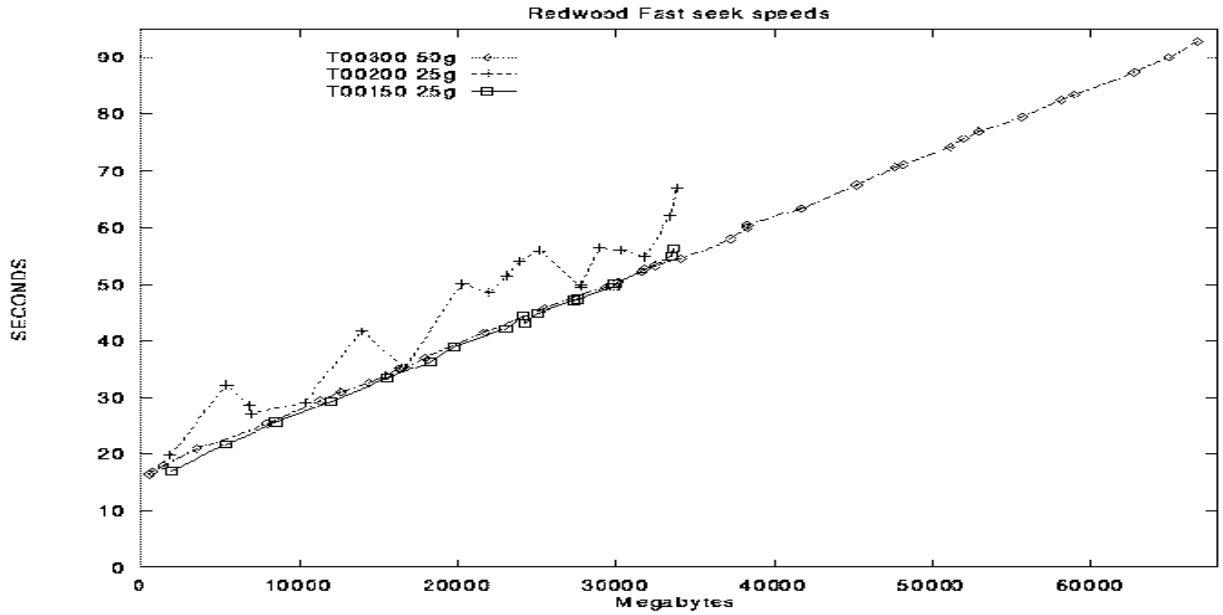

Figure 1: Tape T00200 shows the ILH problem. For small block seeks the Redwood takes over ten seconds.

to gauge the retrieval time. The write transfer rate was between 10-13.5 megabytes per seconds (MB/s) and read rate was 13.6-14.8 MB/s. With DMF, the read and write rates to the Redwood were between 9-10 MB/s. For skip block the rate was 22.5-25.6 MB/s.[1]

For the *locate block id* (seek) the rate is linear except when the ILH problem occurs, as seen in Figure 1, with the sawtooth line of tape T00200. The seek rate is between 700-800 MB/s*. The Redwood takes between 10 to 15 seconds, before the drive can service any seek request. Once a tape is mounted and loaded it takes about 94 seconds to seek to the end of a 50 GB tape and about 57 seconds for a 25 GB tape. For small files or seeking a short distance, the Redwood will perform at its worst. The STK 4490 drive can seek small files or short distances better than the Redwood, as shown in Figure 2.

When the project began, the value to use for the Redwood tape zone size or how the zone size would impact restore time was unknown. Based upon previously stated information the zone size was set to 500 MB. Ignoring the mount time, the worst case seek time for a 50 GB tape is about 114 seconds or on a 25 GB tape is about 77 seconds. For a 3490E tape with a zone size set to 200 MB the worst case seek time is about 112 seconds.

NAS DMF Production Load and Experiences

During the last year, 11.5 TB of DMF data has been written on the NAS CRAY C90, although the current migrated storage is about 2 TB. NAS encourages its customers to use the CRAY home directories for short-term storage and to use NAStore[2] for long-term storage. This makes the tape merge function very important for lowering tape cost and number of total tapes needed. For a recent six month period, the DMF daily average for new data written was 39 GB and data restored was 7 GB. With the current DMF configuration, less than 5 percent of the files used per day need to be reloaded from tape.

Using the advanced tape format over the last several months, there has been a 20 percent improvement in the amount of data stored on a tape. Previously, using the old tape MSP, the tape size was set to 1000 MB, although CRI recommended that the tape size being set to 800 MB. For sites using the 800 MB limit for 3490E tape, it may be possible to have a 50 percent improvement in tape capacity. With the 1000 MB tape size for the old MSP, some tapes would have to be rewritten (a few times a week) because the combination of files at that point in time would not compress well to fit on a tape. A few hours later, though, a different combination of files would be written to tape.

Over the last several months, the old formatted tapes have been merged using the advanced tape format and new migrated data has been written using the advanced tape format. For the 2,057 3490E tapes with the advanced tape format the tape capacity mean is 1207 MB, median is 1182 MB, and the standard deviation is 145. The smallest 3490E tape capacity is 883 MB and the largest is 2502 MB.

DMF management and maintenance uses less than 1 percent of the CPU cycles on the C90. For the operational year, DMF will use about 260 CPU hours, out of over 128,000 CPU hours. This makes it convenient to do computing and archiving on the

[1] Typically *skip block* and *locate block id* (seek) are measured in blocks per second. Using the same units makes it easier to compare the different rates.

[2] Locally developed product for archived storage.

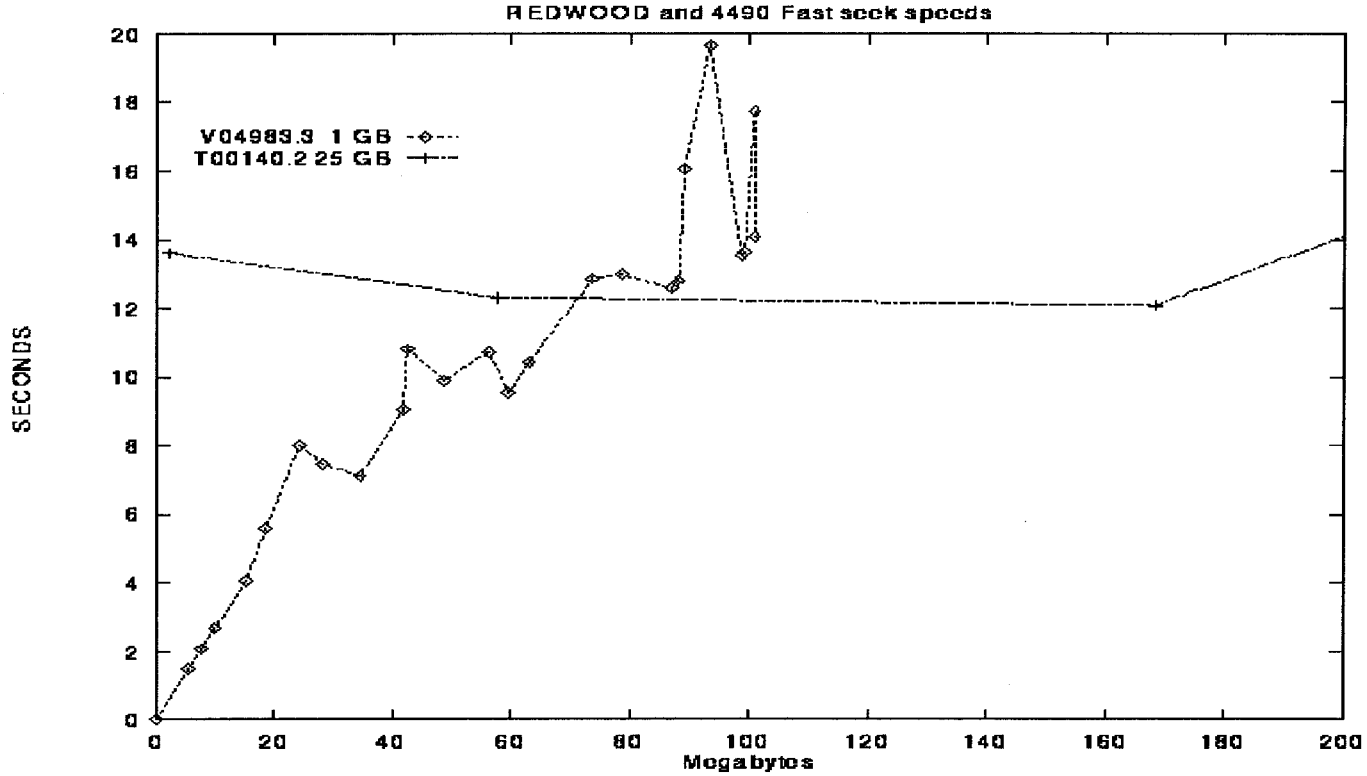

Figure 2: The 4490 tape device seeks short distances better than the Redwood device.

same system. In the future, to save money, other sites may combine the computing and archival platforms into a computational archival storage platform.

DMF Feature Overview

DMF functionality has improved greatly over the years and has become very stable. This section covers different DMF features, including local site experiences and opinions.

DMF is integrated into UNICOS, and having DMF use the native file system allows it to work with disk quotas, the file system daemon, NFS, and DFS. It is possible to export migratable file systems, which provides other hosts direct access to remote storage. DMF depends on the UNICOS tape daemon to provide access to tape devices and automated tape storage units. When new tape technology becomes available, CRI has quickly integrated the new technology into UNICOS.

With CRI's Shared File System (SFS)[3] used between two or more hosts, and DMF in client/server mode, it is possible to provide several DMF servers[4,5]. When the primary server is down on one host, the backup DMF server on another host will automatically process the DMF request.

The Client/Server mode provided in DMF 2.3 allows DMF to hold requests while the DMF databases are compressed or the server is restarted. It is also possible to configure a server for each file system or a combination of file systems. For sites with very large DMF databases and several migratable file systems, DMF databases can be split by file systems. The client/server mode setup for each files system provides a way to dismount a migratable file system without affecting all DMF customers.

DMF is able to write to several different media types. For each media type a separate process (MSP) is used to transfer data to and from the storage media. Several sites have already created an MSP to write to a remote host instead of using locally attached tape drives. In DMF 2.3, this functionality is supported by a new MSP, which uses ftp to transfer DMF files to another network host (UNIX or non-UNIX). It is helpful of CRI to start supporting this feature for all sites to use.

DMF can do simultaneous reads and writes to all tape drives (up to 24) for each tape MSP. A limit can be set to reserve a number of drives for retrieval and to reduce the number for writes. DMF is not limited to how much data it can read or write in a day, but the underlying hardware is. Tapes, disk devices, and channels need to be balanced for any archive system to perform well. If the I/O hardware is slow or unbalanced, then DMF will be slow. Using faster hardware does not make DMF a better product, nor does using slower hardware make DMF an inferior product.

Using DMF 2.2's advanced tape MSP, the tape merge process is automated to reduce the impact on the DMF load. After selecting which tapes to merge, the system checks the DMF load before doing a tape merge set. This MSP takes advantage of locate block id and tape compression, and recognizes end of tape. By doing weekly, merges the tape utilization is about 80 percent for all used 3490E tapes.

Previously, whenever the volume and catalog databases were updated, a copy of them was made. With release 2.2, DMF creates a journal log for each database.

File Percentage	Total Files Count	File Size (MB)	Total File Size (GB)	Total Size Percentage
1	1577	180	634	27
3	4733	90	1011	43
5	7889	63	1242	53
10	15778	34	1596	69
20	31556	14	1942	83
30	47334	7	2092	90
50	78891	3	2234	96
100	157783	0	2326	100

Figure 3: Profile of NAS's migrated files.

Using ***dmsnap,*** it is possible to backup all the DMF databases while DMF is in production. This backup process takes just a few minutes and the DMF requests are held until the backup completes.

DMF has several tools to help verify the integrity of the file system, DMF, volume (tape), and the cat (offline file) databases. ***Dmaudit*** checks the consistency between the file system and the DMF databases. ***Dmatvfy***, ***dmvoladm***, and ***dmcatadm*** check all the tape databases. These tools are helpful in finding and correcting problems. For DMF tapes with write errors, the tools ***dmatread*** and ***dmatsnf*** help read the tape or verify tape integrity.

Since the release of DMF 2.1, ***dmselect*** and ***dmmove*** provide a way to move or copy migrated data to one or several media types. This enables a site to easily migrate data from expensive to less expensive storage. The criteria used by ***dmselect*** is size, age, owner, file system, and MSP. If this is not enough criteria, simply provide the path names of the files to be written to another media type.

DMF creates daily logs used by ***dmastat*** to report usage statistics for migration and recall activity. This helps sites to know the migration load of the system.

DMF provides a very fast tool, ***dmhit,*** to scan the file system to find files to migrate. When a file system is full, releasing data blocks needs to be done quickly, but before this can be done the files need to be migrated.

The hard delete process removes DMF database entries for files that have been deleted prior to a given period of time. Previously, the hard delete process took several hours, but now it takes less than 15 minutes and less than one CPU minute, with just one command, ***dmhdelete***.

Using ***dmcopy,*** it is possible to read a piece of a migrated file, an advantage for sites with very large files. DMF and UNICOS have supported files greater than 2 GBs for a long time.

DMF is very flexible; below are the most useful options:

- Number of copies to write offline for the DMF file.
- The minimum size of the file to be migrated.
- For each MSP the minimum number of restore processes and the maximum number of write processes (up to 24) can be set.
- The archive priority is site configurable by age and/or size. For more flexibility, CRI provides with DMF's binary release the function, *dma_wt.c*, for a site to modify how the archive priority is set.
- The user's *archmed field* in the udb (UNICOS password file) can be set to select the MSP(s).

In the next release (DMF 2.4), the functionality to create dual state files is planned to be provided with the utility ***dmmigall***.[2] For sites with large file systems, this could greatly reduce system dumps. NAS had at one time a 130 GB file system that took more than 12 hours to dump and used over 65 tapes. With NAS's **dmmig** (**dmmigall**), the dump time was reduced to under 2 hours and less than 8 tapes.[2]

DMF 2.3 Beta

NAS has developed several different DMF tools, such as ***dmsnap***[2], to backup the DMF databases while DMF is up, and ***dmmig***[2] to create dual state files. ***Dmsnap*** functionality was included in DMF 2.2, and ***dmmig*** functionality will be included in DMF 2.4, with the CRI utility ***dmmigall***. NAS discussed with CRI the possibility off providing the functionality to compress the DMF databases without taking DMF out of service. CRI understood the need, but would not commit to a release. Later. CRI told NAS that the next release would have the functionality to compress the DMF databases while DMF requests were held. NAS wanted to know when it would be released and CRI invited NAS to beta test the next release and after reviewing the beta test agreement, NAS agreed.

For the beta agreement, CRI wanted to have a DMF contact on site for the first two weeks. The first week was Dinesh Helapitage (developer) and the second week was Jim Zimmer (support). NAS had two test periods with the new version. Both went surprisingly well and the new version of DMF only had minor problems, which Dinesh had fixed by the following day. NAS started DMF 2.3 in a production load early in the week and had no problems until the weekend. On the weekend DMF hung. DMF was restarted several times, and soon after each restart DMF would hang again. After looking at the logs, it appeared that the hard delete process was related to the problem. After stopping the hard delete process, DMF returned to normal. After several weeks of not doing a hard delete, CRI informed NAS that it would be a while before a fix was available and suggested switching to standalone DMF of the same release. In this configuration, the hard delete process finished normally.

NAS stayed with this beta release in stand-alone mode until the official release (2.3.1) was available. With the official release in client/server mode there has been no problems with the hard deletes.

DMF Database Compression

Prior to DMF 2.3, it would take between 2-4 hours of dedicated system time to manually compress the DMF databases.

Much time was spent doing other system activities and switching between dedicated mode (i.e.; check pointing all jobs and disabling user accounts) then back to multi-user mode. With the DMF client server model, the server can be taken down without rejecting DMF requests and can stay in production.

The DMF database compression wall-clock time can be reduced by using an ldcached file system, preallocating the disk space for databases using ***setf,*** and parallelizing the process to compress all the databases. Doing all this reduces the total time to compress all the databases to less than 4 minutes. Old DMF sites are probably already doing this, but this process should be reviewed at new DMF sites.

Reduce 3490E Tapes by 90 Percent

Since secondary copies are made for recoverability and not speed, the secondary copies can be written to the Redwood drives, thus reducing the number of tapes by about 50 percent. It has been shown that the Redwood drives do not handle small (< 64 MB) files well. Files greater than 64 MB can be written to the Redwood tapes reducing the number of 3490E tapes by another 25 percent (Files greater than 64MB are about 50 percent of the primary 3490E tapes, as see Figure 3). Another 15 percent of the data can be saved if all migrated data older than six months are migrated to the Redwood tapes. NAS customers like to have their most recent files available quickly, but this expectation changes with older files. By having the DMF environment configured with the above settings, 90 percent of the 3490E tapes could be reduced. For sites with large data files, the Redwood could replace the STK 4490. These estimates are based on existing NAS data and each site's data mileage may vary.

To reduce the 3490E tapes by 90 percent, the old files and large files can be moved to the D3 tapes using ***dmselect*** and ***dmmove***. Choosing a small number of DMF requests to process at a time (500) for ***dmmove*** can lessen the impact on the system. For NAS this process takes several months, but once done, a weekly procedure takes a couple hours to move the week's data from 3490E tapes to D3 tapes.

It is better to write for the large files directly to the D3 tapes, instead of coping or moving from the 3490E tapes. To write large migrate files to the D3 tapes the DMF function *dmmfunc.c* (This is also included with the binary DMF release.) can be modified to select the appropriate MSP based on size. CRI should provide several items within the DMF configuration file to select the MSP, in order to implement a site's storage policies. Some of these items could be: size, file system, age, file usage, owner, group, and acid.

Suggestions

For new hardware devices. CRI should provide (possibly on the world wide web) performance numbers, reasons for default settings, and conclusions. This would help customers know if the hardware is performing as expected and be aware of known problems.

The DMF manual functionally shows how to do a database compression, but it also needs to show ways to reduce the wallclock time. UNICOS has several tools to help optimize procedures and should include these tools in example procedures within all administration guides.

DMF needs to be able to limit the impact a single user has on the whole DMF system. If a single user requests 10,000 offline files, then all other users' DMF requests will wait until all 10,000 files are restored. Even customers retrieving only one file, will wait a long time. The only way to solve this problem is by educating the offending customer to limit the number of simultaneous DMF restores.

Conclusion

Both CRI and STK have done a good job during the NAS beta tests. The STK Redwood transfer rates were superb for large files, but for small files it is better to use the STK 4490 or 9490 devices. With the current DMF release, a site can hold DMF requests while doing a DMF database compression. DMF has grown to be efficient, reliable, and flexible. By writing migrated data to the most appropriate device (Redwood and 4490), the 3490E tapes can be reduced by 90 percent. With less than one percent overhead for DMF, other sites may also be able to combine both archiving and computing into a single platform. Locally developed DMF tools and ideas have been incorporated into DMF to provide a better solution for everyone. By attentively listening to their customers suggestions, both CRI and STK are able to continue to provide added value for their Commercial Off The Self (COTS) solutions.

Acknowledgments

This work was performed by Sterling Software at the Numerical Aerodynamic Simulation Facility (Moffett Field, CA 94035-1000) under NASA Contract NAS2-13619.

References

[1] DMF Administrator's Guide, SG-2135 version 2.2 and 2.3, section Media concepts, pp 37-39.

[2] N. Cardo, "Improving Recoverability by Utilizing DMF", *1995 Spring CUG Proceedings*, Denver, CO, March 1995, pp. 116-119

[3] UNICOS Shared File System (SFS) Administration, SG-2114 8.0.4

[4] S. Lord, "DFS/SFS Benchmarks", CRI Report, February 24 1995

[5] K. C. Matthews, "Implementing a Shared File System on a HIPPI Disk Array" *Fourteenth IEEE Symposium on MASS Storage Systems*, Monterey, CA, September 1995, pp. 77-88

[6] R. Albers, et al., "CRI RAID", *1995 Spring CUG Proceedings*, Denver, CO, March 1995, pp. 116-119