The Rebirth of DMF on IRIX

Alan K. Powers, Sterling Software Inc., Numerical Aerospace Simulation Facility, NASA Ames Research Center, M/S 258-6, Moffett Field, CA 94035-1000, USA

ABSTRACT: *The Numerical Aerospace Simulation Facility (NAS) at NASA Ames Research Center (AMES) is in the process of testing DMF 2.6.1.4 on a Power Challenge XL and an Origin 2000. These systems are connected to a STK 4400 silo using SCSI STK 9490 tape drives. A list of new features and differences will be covered. A simple benchmark will be used to compare the performance of IRIX DMF and UNICOS DMF.*

NAS Mission

The National Aeronautics and Space Administration (NASA) created the Numerical Aerospace Simulation (NAS) Facility to focus resources on solving critical problems in aeroscience, space technology and related applications by utilizing the power of the most advanced supercomputers available. The mission of NAS is to ensure continuing leadership in Computational Fluid Dynamics (CFD) and related computational aerospace disciplines by:

- acting as a pathfinder in advanced, large-scale computational capability through systematic incorporation of state-of-the-art improvements in computer hardware and software technologies;
- providing a national computational capability, available to NASA, DOD, industry, other government agencies and universities, as a necessary element in ensuring continuing leadership in computational fluid dynamics and related computational aerospace disciplines;
- creating a strong research tool for the NASA Office of Aeronautics.

Production DMF Platforms

Currently the Data Migration Facility (DMF) is on three CRAYs at NAS. They are two C90s and a J90. The C90s have 16 and 8 CPUs, respectively, with 8 and 2 gigabytes (GB) of memory, respectively. The C90s each have 4 tape channel adapters (BMX) connected to four cross-coupled StorageTek (STK) control-units with 32 - 4490 tape transports inside four Library Storage Modules (LSM, aka SILO).

For the large C90, DMF manages four user home file systems of 60 GB each and a staff file system of 22 GB, for a total of 1.56 million files. This system has been using DMF for over 4 years and has offline storage of 2,300 GB. The daily tape traffic is about 39 GB a day.

For the smaller C90, DMF manages four user home file systems of 16 GB each and a 74 GB file system for a total of 1.5 million files. This system has been using DMF for over 5 years and has offline storage of 4,800 GB. The daily tape traffic is about 33 GB a day.

The J90 has 12 CPUs and 4 GB of memory with two SCSI channels connected to four STK 9490 tape drives inside an LSM. DMF manages four user home file systems of 36 GB and a staff file system of 10 GB, for a total of 435,000 files. This system has been using DMF for about a year and has offline storage of 550 GB with a net growth of 8 GB per week. The daily tape traffic is about 10 GB a day.

All the DMF hosts at NAS have similar tape utilization that varies between 75-80 percent, with the average amount of data stored per tape being approximately 1200 MB. DMF is configured to only migrate files greater than 1 MB to tape leaving about 90 percent of the files online. The small files represent less than 3 percent of the data. Each DMF host runs **dmmigall** three times a day to migrate new data older than 1 hour, leaving the data blocks of files on disk (dual state files). A **dump** option (-a) can be used to skip dumping the data blocks of DMF files and only dump the inodes of these files and non-DMF files. These full dumps use only 6 tapes versus the usual 50+ tapes. All the users have the same disk quota: 10 GB of data (online plus offline storage) and a

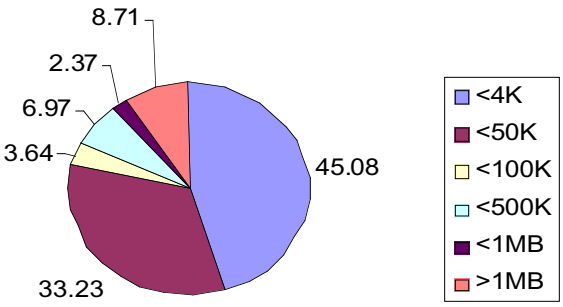
maximum of 10,000 files for each system. These limits help encourage users to transfer long term data to NAS's archive hosts and only leave short-term data on the CRAYs.

Web Accessible Metrics

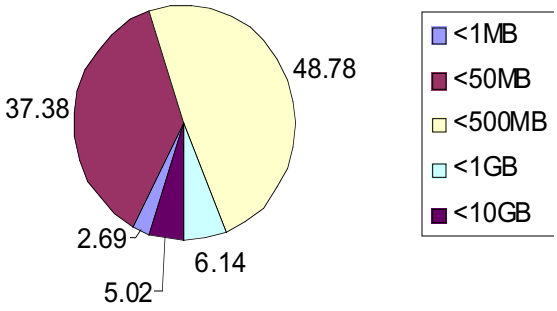
Various online metrics are maintained for each of the CRAYs. These metrics dynamically generate plots for batch job, file space, system, network and hardware performance monitor statistics (MegaFlops). The file space statistics are gathered for each of the CRAYs and below are a couple of typical plots. Less than 9 percent of files are greater than 1 MB represent 97 percent of the data and the other 91 percent of the files represents less than 3 percent of the total data for the large C90. The online metrics are located at:

science.nas.nasa.gov/Groups/HSP/metrics.html

Percentage of Files by File Size on Vn's DMF File Systems



Percentage of Storage by File Size on Vn's DMF File Systems



Testbed DMF Platforms

The first SGI host used for the testing was an Origin 2000 with 64 CPUs and 16 GB of memory running IRIX 6.4. The test file system was approximately 10 GB. There were several problems with accessing the tape drives and the system needed to be rebooted often to clear the problems. This created several complaints from the NAS users pertaining to the down time.

Because of this, a PowerChallenge XL with 4 CPUs, 1GB of memory and two SCSI channels connected to 4 STK 9490 drives was dedicated to testing DMF. The XL had IRIX 6.2 for the operating system. With this system, DMF was successfully able to read and write tapes.

Later a small O2000 with 4 CPUs, 1 GB of memory and IRIX 6.4 became available for dedicated testing. After adding the latest required patches for DMF and during the reboot, the tape devices were not sensed and hence were not added to the device tree. It took months for SGI to provide a work-around for the problem. Once this platform was moved into production, DMF testing stopped on this platform.

Another O2000 with 8 CPUs and 2 GB of memory running IRIX 6.5SE is currently being used for testing. The system has 2 SCSI channels connected to 4 STK 9490 tape drives. The test file system is about 15 GB using two SCSI disks with striped partitions.

DMF IRIX Evaluation

As of this writing SGI is offering a **free** evaluation version of DMF. The latest evaluation version of DMF can be downloaded from the web site:
www.CRAY.com/products/software/storage/dmf/index.html
The compressed tar file is less than 10 megabytes (MB). Each site needs to fill out an online form for a contact person to receive a 30-day temporary license via email. Once the online form is filled out the key should arrive within 5 minutes.

DMF Planning Process

For an existing DMF site, it is important to understand the amount of daily traffic (tape and ftp) and typical file usage by size and age per user and application. For DMF to benefit a site, the users should still be able to be productive. Users mostly access or modify small text or source files. These files should be the last to be migrated, or if there is enough disk space they should never be migrated. Larger, less used, and older files should be candidates for migration. A site should plan how long new and recalled data is to stay online. NAS goal is to have new and recalled data stay online for 4 days. For example with the large C90, there is over 260 GB of online storage with daily tape transfers of 39 GB. On average, data stays online about 6 days. With peak days of 75 GB of daily tape transfers, data will stay online just under 4 days. But each site is different: the IRIX version of DMF can easily be configured with almost any migration policy.

New sites need to estimate daily tape traffic and typical file access patterns. It is difficult to know file access patterns just by surveying the customers. Unless the customers are good at estimating their data usage, most likely the estimates will be off by a large margin. Set some goals such as minimal impact to the users, new and recalled data stays online for some period of time, and frequently used small files stay online forever. Whatever the goals are, be willing to change them within a few weeks to a month. After analyzing the real access patterns of the users' data use, make the necessary adjustments.

Choosing how large each file system should be will vary for each site. A goal is to have file system dumps complete quickly. The longer the time period, the more likely (on UNICOS) inodes or data have changed. Of course, in recovering from the loss of a file system, the longer the dump the longer the restore. The NAS goal is to have a dump complete for a file system in less than 4 hours.

Typically the DMF databases (dmf, tape, cat) are located in /usr/dm and the journal files should be located on a separate file system and disk. For DMF configured with two tape media specific processes (MSP), use 2000 bytes multiplied by the number of DMF files to estimate the initial size of /usr/dm. This estimate should provide some room to grow. The initial size of the journal file system can be the same as /usr/dm. The journal file system should have the latest backup of the database (from **dmsnap**) and about a week of journal logs.

DMF IRIX Installation

The install process is best covered in the *README* file included on the DMF download web page. This reviews the patches to install for the various operating system versions, mount options for DMF file systems and how to configure tapes for IRIX. The patches should be installed before DMF is installed. SGI customers with a software maintenance user id and password can downloaded the patches from web site:

support.sgi.com/index.html

The *README* file assumes basic knowledge of how to install products for IRIX. The process would be better documentation if these few simple steps were included to support CRAY administrators with little IRIX experience. The SGI commands **inst** or **swmgr** can be used to install products.

Assuming the patches and SCSI drivers are correct, the SCSI tape drives can be connected to a SCSI channel and the system rebooted. Once rebooted the devices are automatically sensed and added to the device (/dev) tree. To display the tape devices the **hinv** command can be used as shown below for the STK 9490 tape drives.

```
% hinv
…
Integral SCSI controller 5: Version
QL1040B (rev. 2), differential
  Tape drive: unit 6 on SCSI controller
5: STK 9490
  Tape drive: unit 7 on SCSI controller
5: STK 9490
```

The tape subsystem for DMF will be changing in the next several months to use the tape management facility (TMF). TMF is a port of the UNICOS tape subsystem.

DMF Changes

UNICOS DMF uses custom system calls to migrate and recall data. IRIX DMF uses a newly developed standard set of system calls, the Data Management Application Programming Interface (DMAPI). For more information on DMAPI, Open Group has a book called System Management: Data Storage Management (XDSM) API. Free access to an online version of this book is at: www.opengroup.org/publications/catalog/c429.htm Currently, SGI does not provide any documentation of their DMAPI extensions.

The IRIX DMF implementation is modeled after the IEEE storage model. With UNICOS, each DMF file has a machine id and a file id. With IRIX, each DMF file has a file handle (fhandle) and a unique bit file identifier (bfid). The installation directory has changed from */usr/lib/dm* to */usr/dmf/dmbase/VERSION*. The current version is 2.6.1.4. There is also a directory */etc/dmf* with a symbolic link of *dmbase* to the active installation directory. The installation directory has a *bin* subdirectory for user commands and the *etc* subdirectory for system commands. In addition the man pages are kept in sub-directory *man* and the DMF configuration file is located in sub-directory *host/HOSTNAME/dmf_config.* With the changes in location, it is easier to test new versions of DMF and to switch back to the production version.

Previously DMF used the message daemon to provide important information, which could be monitored by the **oper** command under UNICOS. Under IRIX, DMF sends important messages to **syslogd** which can be monitored by IRIX command **sysmon**.

CRAY provides online manuals via the web interface called *dynaweb*. SGI provides online manuals via an X interface called **insight** and a web version using *supportfolio.*

New User Commands

There are three new user commands, **dmfind**, **dmls**, and **dmattr**. As one would expect, **dmfind** and **dmls** are like **find** and **ls**, but have DMF related options. The command **dmfind** has options to select files by *bfid*, *fhandle* and *DMF state.* The command **dmls** list files with DMF state information. The commands **dmfind** and **dmls** will likely be integrated into IRIX's **find** and **ls**. The command **dmattr** displays a list of attributes about DMF files. Using the ***–a*** option provides a way to select the attributes to display and the order to display them.

```
% dmattr -a owner,size,state,path JuNk.*
6012 134217728 DUL JuNk.128.2
6012 134217728 MIG JuNk.128.3
6012 134217728 MIG JuNk.128.4
6012 2097152 DUL JuNk.2.1
```

System Commands and Features

There have been several changes from the UNICOS version of DMF to the IRIX version of DMF in the systems commands and features, some are new or renamed, and some have been replaced or are missing.

New System Commands and Features

The command to help install different versions of DMF is **dmmaint.** This is a simple X interface with different buttons to install, update license, configure, activate and check for errors. Having a starting place for DMF is a good idea however, this would be more useful if it were integrated better.

For sites wanting to convert from another hierarchical storage manager (HSM) to DMF, SGI has provided a set of tools (**dmcapture, dmloadfs, dmftpmsp**) to make this job simple. These commands are in IRIX DMF 2.6 and UNICOS DMF 2.5.5. The command **dmcapture** is provided with C source as a reference. This code is to run on the old HSM system to create a *capture* file of each file inode information. With a **dmcapture** option (-s) the data for files under a specified size is also stored in the *capture* file. The **dmloadfs** command is executed on the DMF host to read the *capture* file to create the files and inodes from the old HSM. The DMF host uses the media specific process **dmftpmsp** to transfer the data blocks from the old HSM. Using **dmselect/dmmove** a process can be setup to automatically convert from the old HSM to DMF.

To create text copies of the DMF transaction journal logs use the command **dmdumpj**.

The DMF configuration file includes three new policies; *file system management*, *file migration/free* and *msp selection*. The *file system management* and *file migration/free* policies are defined together for a DMF file system. These policies can be the same or different for each file system.

The *file system management* policy defines the high and low water marks of a file system. These determine when to start archiving and when to stop releasing data blocks for a DMF file system.

The *file migration/free* policy determines order by a site's weighting factor for files to be migrated or dual state file data blocks to release. In the past, only size and age of a file multiplied by a constant was used to determine the weighting factor. Now user and group id can be used to determine the weighting factor. In addition to this, a range or set of values can be used with logical AND and OR.

In the complex example, an age weight factor is for files less than 10 days old, greater than 1 GB in size and owned by one of the user ids 10, 82-110, and 200.

```
AGE_WEIGHT  1  0.01  when age < 10 and
space > 1g and  uid in (10 82-110 200)
```

The *msp selection* policy determines which, if any, MSP to use for archiving. The MSP can be chosen by size, age, uid or gid. Multiple MSPs can be selected.

In the example below, files less than 1 MB are not archived, files greater or equal to 1 MB and less than 2 MB are archived using a ftp MSP and files greater or equal to 2 MB are archived to two separate tape MSPs.

```
SELECT_MSP none     when space < 1048576
SELECT_MSP ftp_c    when space < 2097152
SELECT_MSP ct1 ct2 when space >= 2097152
```

Renamed System Commands

The command **dmhit** has been replaced by a better, more flexible command, **dmscanfs**. This command can display the different DMF attributes, inode information, DMF weight (calculated from the rules in the DMF configuration file), and which MSP to use for archiving all the files in the DMF file system. Using the ***–o*** option provides a way to select the field names to display and the order. The ***–r*** option provides a slower method of displaying the information but also includes the pathname.

```
% dmscanfs -r -o bfid,state,space,age,weight,msps
/dmf
…
0 REG 8388608 0.011667 0.000000 cart1,cart2 /dmf/X
352070f40000000000000293 DUL 2097152 1.541111
0.000000 cart1,cart2 /dmf/J28
352070f40000000000000294 DUL 2097152 1.541111
0.000000 cart1,cart2 /dmf/J29
…
```

The command **dmdbase** has been replaced by the easier and more flexible command **dmdadm.** This command provides directives to administer the DMF database. These directives are similar to the **dmvoladm** (tape database) and the **dmcatadm** (tape catalog records database) directives.

The command **dmmigall** has been renamed to **dmmigrate**. This command scans a DMF file system to select non-DMF files to create dual state files. A dual state file has data blocks both online and offline.

The combined functionality of **fsmon**, **fsdaemon** and **dmmctl** has been replaced by the command **dmfsmon**. It monitors the free space of the DMF file system and when a high water mark is reached it can migrate files or free data blocks of dual state files.

Missing System Commands or Features

Two useful commands missing from this release of DMF are **dmastat** and **dmmspuse**. The command **dmastat** provided usage statistics for migration and recall activity. The command **dmmspuse** provided volume and age statistics of migrated files to generate an ASCII plot.

With this version of DMF there is no client/server support. This was primarily intended for Cray's shared file system, but NAS uses this configuration to support taking down the server without losing existing users' **dmget** requests.

Under IRIX 6.5SE beta, **xfsdump** and **xfsrestore** do not support the option to just dump inodes of DMF files nor do they restore the DMF inodes correctly. But under IRIX 6.4 this feature is supported and in 6.5.1 this feature is scheduled to be supported.

DMF attributes in the users' *udb* entry and the *.keep* file are not supported.

What Your Vendor Never Told You But Should Have

The command **dmdbcheck** was quietly added to DMF 2.5 (I think) with only two lines in the system administrator guide. This command is one of the most important commands in DMF. The only good documentation is in the online man page. This command checks for problems in each of the DMF databases (daemon, tape, cat). This command should be used each time a **dmsnap** is done to validate the database. When recovering from a disaster this command should be run on the most recent snap of the databases and after the relevant journal entries have been applied to the snap of the databases.

Benchmarking DMF

A shell script, archive_bm and C program is used to do the DMF benchmark. The C program is used to create the data files. By default it creates 63 files of the combination of 32 2 MB, 16 8 MB, 8 32 MB, 4 128 MB, 2 512 MB and 1 2048 MB for a total of 4032 MB. The data files have a non-

repeating pattern to try to make file data non-compressible. Once the files are created all the files are migrated to two tape MSPs (two copies) and the data blocks released. The last step is to read all the data back online. Timing in minutes is done for writing to and reading from tape. This process is done at least three times and then averaged. There was no other tape or DMF activity during the test.

There is a critical SPR (711694), for the O2000. It is not possible to write to two tape devices at the same time connected to single SCSI channel.

The current tape support for IRIX does not permit different tape MSPs to share drives as can be done under UNICOS. This is planned to be resolved when UNICOS like port of the tape subsystem is available for IRIX.

Writing to tape, the J90, O2000 and XL are close to the same DMF/tape performance. Reading from tape, J90 has the best DMF/tape performance. This is due to the DMF primary MSP being able to use all the tape drives and bandwidth. The XL also did well in reading from tape. This was not expected because only half the tape drives was being used for reading.

For the test, SCSI channels and STK 9490 tape drives were used and were located in the same LSM. To decode, 1s2t means 1 SCSI channel and 2 tape drives.

System	S/T	Dmput	Dmget
J90	1s2t	34	21
J90	1s4t	27	14
J90	2s2t	32	17
J90	2s4t	23	11
O2000	2s2t	24	23
XL	2s4t	24	12

Support Issues

Our site has tens of millions of dollars of SGI/CRAY hardware, 24x7 maintenance contracts and 3 full time on site analysts. In addition to this, NAS has several memoranda of understanding (MOUs) to facilitate working closely with SGI. But the best part is NAS is right next door to SGI Mountain View headquarters.

During the testing of this product a number of SPRs have been submitted as is expected for a new product. In CRInform there were 78 SPRs created in this last year for the IRIX version of DMF. It is interesting that the architecture type is listed as a SUN or type ‘other’.

For SPRs directly related to DMF, there has been great support. But once a critical problem involves another part of the operating system, the problem can take months to be resolved. Once the right level of person is working on the problem, it can be solved quickly. The problem is getting to the right person. It seems the process is broken when multiple workgroups must work together to solve critical problems.

A typical response is “Are the latest patches installed?” There have been times when the latest patches are installed, something else breaks and the problem is still not fixed.

There might be a number of good reasons why this is happening, but it does not matter. NAS has paid for premium service and does not want to be treated like a workstation customer. If NAS can’t get premium service, then what site can?

Recommendations

For the Vendor

The current process of solving critical problems between workgroups is not working well and needs to be changed so any critical problem can be solved in a timely fashion.

Make the DMF installation process simple enough for an inexperienced administrator to install the product easily.

Make **dmmaint** more useful by being able to import a previous version of the DMF configuration file, edit any DMF/tape configuration file, and have a “Start Here” button to include the steps to do before the activate button.

Have **xfsdump/xfrestore** support dumping and restoring inodes of DMF files.

The tape support system needs to provide capabilities for at least a force dismount, a display of current tape activity, bringing up or down a drive, and being able to sense and add SCSI tape drives to the device tree without a reboot.

For other Customers

When submitting an important problem report, submit to both *crinform* and *escall* to increase the chances of finding the right person to fix the problem. In addition, follow up with phone calls within a short period of time.

The IRIX version of DMF has been enhanced to make it easier and more flexible to configure and use effectively. Customers should download the evaluation version to learn about the new features and install it on a test system. The test system can be done on a SGI workstation with 5-20 GB

of disks and using the *ftpmsp* to migrate files to another host.

Although this version of DMF shows promise and can be used for testing, the tape support needs to be better before it is used for production.

References

Cray DMF Release and Installation Guide for IRIX Systems SG-5299 2.6.1

Cray DMF Recovery and Troubleshooting Guide for IRIX Systems SG-2217 2.6.0

Cray DMF Administrator's Guide for IRIX Systems SG-2216 2.6.1

Cray DMF Administrator's Guide SG-2135 2.5

Acknowledgments

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

The author can be reached at powers@nas.nasa.gov and my online CUG papers can be accessed on the Web at www.nas.nasa.gov/~powers.

Additional Reading

[1] FAQ for comp.arch.storage part 1 www.faqs.org/faqs/arch-storage/part1/preamble.html

[2] FAQ for comp.arch.storage part 2 www.faqs.org/faqs/arch-storage/part2/preamble.html

[3] IEEE Storage System Standards Working Group www.ssswg.org