WHITE PAPER

sgi®

Managing Large Data Securely and Cost Effectively with SGI DMF 5

By David Honey (SGI)

SGI's DMF 5 - Advanced Tiered Virtualization

DMF is an enterprise-class Hierarchical Storage Management (HSM) software tool with twenty years of real-world use and hundreds of installations. DMF enables organizations to cost-effectively manage storage needs without limiting user access to data. DMF creates and automatically manages a tiered virtual storage environment that can significantly reduce equipment and operating costs, improve service levels and lower risks.

Based on site-specific criteria, DMF continuously monitors and automatically migrates data between different storage tiers, each with different cost and performance characteristics. In general, the most critical or timely data migrates to higher performance, more expensive storage media, while less critical or timely data is automatically migrated to less expensive, lower performance storage media. Data always appears to be online to end users and applications regardless of its actual storage location.

DMF's Active Copy feature allows multiple copies of data to be created during the migration process. Because copies of critical data already exist, DMF can significantly reduce the time it takes to perform backups since only file system and DMF database information need to be copied during the backup process.

Unlike many other virtual storage managers, DMF can integrate "green" storage technologies, such as automated tape libraries, into its virtual storage pool. According to The Clipper Group, a Boston-based analyst firm, tape can be up to 23 times less expensive than SATA disks and use up to 290 times less energy. Additionally, when combined with MAID technology such as SGI® COPAN™ 400, users are able to achieve disk-based performance while also maıntaining space and energy savings.

DMF's industry-leading features include automatic data migration, automatic data recall, file and automatic free space management, volume level migration policies, custom migration policies through user-defined plug-ins, accelerated data access, partial file migration and recall, Active Copy, GUI-based administration and tape library management.

DMF can manage Petabytes of data, billions of files and run storage devices at their maximum rated speed. DMF operates in the background so there is no interruption or degradation of service to end users and applications. DMF is a fifth generation software product owned and developed by SGI.

Benefits

Usability

DMF software virtualizes disk and tape infrastructure to present a 'normal' volume with unlimited capacity to users. A DMF managed volume contains all of the directories and file system metadata of a typical file system and imposes no software requirements or constraints on users other than first access latency.

DMF useability features:

- Transparent user access – no special steps, applications or data changes are required
- Automatic space management (file systems never fill up because DMF detects file system thresholds and removes least recently accessed data from primary storage while retaining file metadata)
- Accelerates backups by not backing up files that have not changed
- Automatic tape media consolidation
- Browser based GUI
- Powerful utilities to aid data management and storage updates
- Partial file support
  - Users working with very large files can select interesting regions to retain online while migrating the bulk of a file to lower tiers
  - Users can read online regions while the 'backend' of the file is being recalled from tape

Reliability

Data reliability is not just a day-to-day issue for active archives, but also year-to-year and even decade-to-decade. DMF software has been continuously improved to provide our users continuity of data access as data volumes have mushroomed. Customers who have upgraded servers, disk and tape storage multiple times can attest to being able to access files that still contain important data that was created 16 years ago.

DMF customers are quoted as saying that they would not trust their data to any other product. SGI can arrange references that customers can contact independently who would be happy to testify to DMF reliability.

Features built into DMF that make it resilient to unscheduled interruptions include:

- Automated copy creation and management
- Support for 1 – 63 copies of data on RAID protected storage and mixed tape formats
- Positive verifiable data integrity
- Checksum calculations on every tape block written
- Checksum comparison on every tape block read
- Journaling database supporting Atomic, Consistent, Isolated and Durable transactions
- Automated (daily) auditing
- Automated (daily) tape drive error reporting
- High Availability (automatic failover to standby server)
- Tools for managing tape pools
- Streamed data reduces wear and tear on drive mechanics and tape media and increases performance
- Automated media error detection and tape substitution
- Low level media recovery tools

Scalability

DMF is installed at many worldwide sites that demonstrate the scalability of DMF in terms of the number of files (in excess of 200 million at a film production customer), the number of bytes transferred per day (over 120TB per day at a climate customer) and the volume of data that can easily be managed (over 40PB for one government customer). DMF is used at sites which migrate hundreds of thousands of files per day, file sizes range from KB to TB and can comprise any data format.

DMF scalability is superior to that of other HSMs due to the following factors:

- Scalable CPU, memory and IO platform (SGI® Altix®, distributed I/O architecture)
- High capacity file systems – XFS and Clustered XFS
- Efficient software, low CPU overhead, optimised for I/O
- Support for high tape drive and slot count tape libraries greater than 28,000 slots.
- Support for the latest high performance tape drives; T10000B, T9840D, TS1130, LTO4 and Super-AIT2
- Multi threaded multi process server architecture. I/O child processes are independent of DMF server processes in monolithic and distributed implementations
- Advanced Features include:
    - Disk Cache tier
    - Native commands for Linux, Solaris, IRIX ,Mac OS-X, CXFS and NFS clients
    - Partial File Migration and Retrieval
    - File tagging and customized policies
    - Clustered file system (CXFS) support

Performance

DMF performance is demonstrably superior to other HSMs due to the following factors:

- High performance server platform (SGI Altix)
- Clustered I/O architecture (based on CXFS)
- High performance journaling file system – (CXFS -- tested to 45GB/s)
- SGI-developed hardware device drivers
- Support for high performance disk, MAID, tape and robotic devices
- High performance algorithms move large amounts of data efficiently and ensure tape devices stream
- Files are migrated in parallel to saturate channel bandwidth
- Migrated files are stacked to keep tape buffers full
- Secondary disk cache
- Fast tape positioning
- Highly optimized tape movement scheduler
- High integrity software architecture

Flexibility

The DMF software architecture consists of a daemon and one or more Library Servers (LSs) or Media Specific Processes (MSP) and a mounting service. In addition to tape libraries, DMF can manage local tertiary stores (disk MSP and disk cache MSP), and remote tertiary stores (FTP MSP) which allows DMF to manage data across Wide Area Networks (WAN) and between independent DMF servers. The FTP MSP has been exploited by the University of Queensland's Centre for Magnetic Resonance, Earth Systems Sciences Computational Centre and James Cook University to provide an uncompromised level of data security over campus LAN and WAN links.

In both Clustered XFS and NFS shared file system environments, DMF provides distributed data management commands for execution on cluster nodes other than the DMF server. A client API is also provided for users to create their own site-specific data management policies.

The DMF software architecture allows for multiple Volume Groups of different media type and generation, multiple Library Servers (one per tape library), multiple tape drive groups (typically one per drive type) and an unlimited number of policies including customer developed policies created via DMF's API or SOAP (Simple Object Access Protocol) web service interface. File tagging allows administrators to uniquely identify nearly 4.3 billion data classes by any criteria, together with site specific policies gives administrators the ability to alter the DMF server's behavior in order to achieve any conceivable data management process.

Quality

DMF is fully qualified and tested by SGI. SGI conducts integrated QA testing of InfiniteStorage solutions; servers, kernel enhancements, storage, switches, volume manager, drivers, HSM, HA clustering and file systems and utilities on a unified environment prior to release.

SGI Manufacturing operates an ISO9000 accredited quality system.

Openness

DMF runs on SGI's XFS or Clustered XFS file systems on a wide range of storage hardware. XFS is the foundation of SGI shared file system (CXFS). DMF managed file systems may be accessed using common standard IP protocols including NFS v2, 3 and 4, CIFS, FTP, SCP and HTTP.

In a Fibre Channel or InfiniBand SAN environment, DMF file systems can be accessed directly from SAN connected nodes running Linux, Windows and Mac OSX.

DMF supports the following tape formats:

StorageTek T9840A, B, C & D, T9940A & T9940B, T10000A & T10000B IBM 3590, 3950E, 3590H, 3592J1A, 3592E05, 3592E06 LTO, LTO2, LTO3 , LTO4 and LTO5. AIT1/AIT2/AIT3 & Super-AIT1 and Super-AIT2 DLT 2000/4000/7000/8000 SDLT 220/320/600

DMF supports a range of tape libraries.

- All StorageTek libraries controlled by the ACSLS interface, release 5.1 or later
- StorageTek L40, L80, L180, L700, L700e, L1400, SL500, SL3000 and SL8500
- StorageTek 9710, 9714, 9730, and 9740
- IBM 3584
- SpectraLogic T50e, T120, T200, T380, T680, T950 and TFinity
- All ADIC libraries that use the DAS interface including the AML-series, Scalar 1000, Scalar 10K, and dual-aisle Scalar 10K
- ADIC Scalar 24/100/1000/10K
- ADIC i500/i2000
- Sony DMS-B35 and DMS-PSC "PetaSite" libraries, and Sony CSM-200, CSM-100, and CSM-60 libraries

Lower Cost and Lower Risk

DMF can dramatically lower equipment and operating costs by integrating automated tape libraries into a Tiered Virtual Storage Architecture and automating the migration and recall of data. File and volume level migration policies permit fine-grained management and storage optimization.

Future storage expenditures are also reduced since the cost of tape library expansion is significantly lower than disk.

Studies by the Clipper Group, an independent enterprise storage consulting company, detail the comparative costs of acquisition and operational costs of disk based and tape based storage. See them at

http://www.clipper.com/research/TCG2008056.pdf
http://www.clipper.com/research/TCG2006046.pdf

Studies by IT industry researchers (Gartner and IDC) show that data administration overheads for decentralized storage far exceed those of centralised architectures. DMF lends itself well to a centralized architecture and enables customers to support user's ongoing acquisition and storage needs for ever increasing data volumes with sustainable costs.

Because daily auditing automatically reports exceptions, it is normal for a part time operator to manage multiple Petabytes of data in a DMF archive.

By minimizing the need for operator intervention, DMF can lower the risk of operator error and improve service levels when dealing with conventionally "archived" data. With DMF the concept of a conventional "archive," virtually disappears since all data is always directly accessible by end users and applications.

Green Energy Efficiency

From the Clipper Group Report #TCG2006046 dated June 4, 2006, "Tape and Disk Costs – What It Really Costs to Power the Devices":

> "Our finding is that for long-term storage over our five-year study period, the cost of (SATA) disk is about 23 times that of (LTO4) tape, while the cost of energy for (SATA) disk is about 290 times that of (LTO4) tape".

Two hundred and ninety times is a staggering differential and results largely from the minimal power and cooling requirements of tape libraries compared with RAID "the initial deployment of the library, with two drives, will consume about 1,150 kWh during the first year, significantly less than the energy consumed to make your morning toast and coffee for a year".

In the Clipper study, the RAID system using 750GB SATA disk drives was estimated to consume between 96,360kWh and 131,400kWh per annum. This system was capped at 245TB though if repeated today, the same system would have provided greater capacity by using 1TB SATA drives with a similar power footprint.

The tape library in the study scaled from 75TB to 2PB with the addition of expansion frames and tape drives rising in power consumption to 2,008kWh per annum in year 5.

SGI would point out that the Clipper study is modeled on a backup application and that HSM systems generally require a higher number of tape drives and require a portion of RAID storage. Power consumption of a single LTO4 tape drive is similar to that of three disk drives. A 'real world' HSM tape library would use power in the same order of magnitude as shown in the study.

Features

Standards Compliance

DMF is implemented in user space and SGI's implementation of the industry standard specification for data management applications to interface with the OS kernel and file systems. Open Group's Data Storage Management (XDSM) API is described in full at http://www.opengroup.org/pubs/catalog/c429.htm.

SGI's implementation is known as the Data Management API (DMAPI).

Layered Application Architecture

Because DMF is implemented in user space it provides a more reliable platform than some HSMs that are tightly integrated with the kernel.

Integration with the kernel causes constraints for applications brought about by OS updates and patches which have the potential to negatively impact application stability and performance. Integrated applications are typically less stable and more prone to data corruption problems. Storage Management systems integrated tightly with the kernel must also be tested and verified with each release of the operating system.

Because DMF uses the Data Management API, its interface with the file system from user space is totally unaffected by operating system upgrades, changes and patches. By using the Data Management API interface, DMF is also insulated from file system changes.

Integration with Data Protection Products

EMC Networker, Atempo Time Navigator and SGI xfsdump are backup applications that can run on the DMF server for backing up DMF managed file systems. Files that have been migrated to DMF tape are not copied to backup tapes in their entirety as DMF tape copies already exist. Rather, the inode contents (512 Bytes), is backed up making backup and recovery processing of vast data volumes very fast and minimizes the use of server and network resources and tape media.

Integral Database

DMF uses a real time journaling database supporting Atomic, Consistent, Isolated and Durable (ACID) transactions.

The use of a journaling database with atomic, 2-phase commit provides positive, verifiable safety of data. Storage management systems must guarantee that the data they store is safe and verifiable the only way to provide this guarantee is with database technology.

For other HSM's that rely on metadata stored on tape, the administrator is given two options for handling users who move or rename their files:

1) Retain the user created pathnames on the tar file, meaning the move invalidates the tape copies
2) New user created pathnames are translated to something of HSM's choosing.

Because DMF uses a database, file movement and renaming is done transparently to the user without invalidating tape copies or requiring the use of a new filename.

Media Consolidation

As files are removed from an HSM managed file system, gaps develop in tape based data. At some administrator determined threshold, DMF will automatically merge sparse tapes onto fresh media.

HSMs that do not maintain a database must perform multiple file system passes to perform sparse tape merging necessary to defragment tapes that have had deletions. These resource consuming scans are required to calculate the size of all the archive copies on each volume, calculate in-use data and then gather information on total capacity and unused space. This is needed for the recycler to determine how much space is used for migrated copies of data and how much is used for expired archive copies. This restriction makes sparse tape merging cumbersome.

DMF, by using its journaling database, constantly maintains space usage information and file information about all tapes; onsite and offsite, in-library and extra-library. With this information, administrators can manually manage tape consolidation (called tape merging) or use SGI provided utilities. This information allows it to perform multiple, simultaneous tape merges resulting in improved media utilisation. DMF can perform this merging operation using a single drive, if desired. The speed of this process is only restricted by the number of available tape drives.

Integrity Checks

DMF always compares stored CRC checksums in the header and trailer of each tape block when recalling from tape. These checksums are calculated when the tape was written to ensure complete data integrity and guard against media errors. Media checking is done to guarantee the safety of the data and the ability to verify that the data is indeed correct on tape.

Some HSM products use the tar format for tape media which cannot detect media contamination, tar also assumes tape positioning is always accurate which clearly is not valid given tape drives behavior of rewinding after a SCSI reset. If the HSM writes to tape or the file system when a tape is incorrectly positioned the result will be data corruption.

DMF routinely manages files well over the size limit imposed by tar. Files managed by DMF are limited only by XFS, a 64-bit file system with a 9 million Terabytes (theoretical) file size limit.
DMF has full auditing capabilities to ensure that data within the database, on the disk, and on the tape are consistent. By default, an audit check is run daily and administrators notified by email of any potential data, hardware or media issues.

Data Structure Independent

Some HSMs work with archive sets, which are sets of files or directories. Each potential file to be migrated is assigned to an archive set and each archive set uses a pool of tapes. Since files are migrated as a batch, it is highly likely that some of the files will not need to be migrated (due to frequent use, for example), resulting in thrashing, or immediately recovering migrated files. DMF provides much lower granularity with its migration policies because individual files can be migrated. This migration methodology ensures that only those files that need to be migrated are migrated. Files which are frequently used are not arbitrarily forced to migrate.

Disaster Recovery

By storing a tape copy at a secure offsite location together with regular full file system and database backups, customers enjoy a simple and cost effective Disaster Recovery mechanism.

The integral DMF database allows administrators to monitor offsite media and track free space as the number of empty regions on tapes grows over time.

SGI tools manage tapes stored offsite, reporting when they should be returned for consolidation and automatically ejecting full tapes for relocation offsite.

A sample of DMF customers:

Advanced Data Processing Research Institute (India)
Centre for Mathematical Modelling & Computer Simulation (India)
Central Research Institute of Electric Power Industry Materials Science (Japan)
Dept of Environment and Climate Control (NSW)*
Dept of Energy (USA)
Dept of Justice (USA)
DownUnder GeoSolutions*
Drug Enforcement Agency (USA)
Defense Threat Reduction Agency (USA)
Diamantina Institute*
Drzavni HidroMeteoloski Zavod (Croatia)
Environment Canada eResearch South Australia*
EuroNews
FBI (USA)
Ford France
Television Publicite
Fuji Television Network
General Motors
Guanhua Glory AV System Integration Co
Idaho National Laboratory
Institut Francaise du Petrole
Institute for Molecular Bioscience*
Instituto Nacional de Meteorologia (Spain)
INA (French Ntl Institute for Audio & Video)
James Cook University*
Japan Agency for Marine-Earth Science
Jiuxingdianguang Technology
Dept of Defense (USA)
DTU Centre for Biologisk
Deutscher Wetterdienst (Germany)
Ecqualis
Fairfield Industries First Automotive Works (China)
Fleet Numerical Meteorology and Oceanography
Freie Universitat Berlin
Frauhofer Institute for Reliability and Microintegration
GE Aircraft Honeywell International

Indian Metereological Department
Institut für Atmosphärenphysik (Germany)
INA (Croatia, Oil&Gas)
Nemzetkozi Technologiai (Hungary)
Japan Meteorological Agency
Kungliga Tekniska Hogskolan (Sweden)
Norges Teknisk-Naturvitenskapelige Universitet
Nederlandse Omroep Stichting (Dutch National Broadcaster)
Northrop Grumman
Nuclear Research Institute (Czech)
Ohio Supercomputer Center
Open Systems Solutions Incorporated
Opel AG
Pertamina (Indonesian Govt Oil Co.)
Pittsburgh Supercomputing Center
Premier Media Group (Fox Sports)*
Proctor & Gamble
ProSeiben Sat1 Produktion
Qinetiq
Question D'Edition
Queensland Dept Natural Resources and Water*
Queensland Cyber-Infrastructre Foundation*
Queensland University of Technology*
RadioTelevision del Principado de Asturias
Raytheon
Retired Power Telecommunications (Taiwan)
RedBridge IT
Robarts Research Institute,
University of Western Ontario (Medical)
Robert Bosch Krankenhaus
Rutherford Appleton Laboratory (UK)
Sudwestrundfunk
Scripps Oceanographic Institute
The Multimedia Corporation
Texas A&M University
Tohoku University
UCLA Laboratory of Neuro Imaging
University of Chicago
Roshydromet (Russia)
Saarländischer Rundfunk Stichting Academisch Rekencentrum (SARA Holland)
Scripps Oceanographic Institute
SKY Channel*
SONY (USA)
State Oceanic Administration (China)
Tata Motors
Telefonica (Spain)
Technical University of Dresden
TBWA (France Prepress)
Total Oil (France)

University of Cambridge
University of Leicester
University of Melbourne - Howard Florey Institute*
University of Montreal
University of Newcastle Upon Tyne
University of Stuttgart
University of Texas (Austin)
University of Tasmania- TPAC*
University of Tokyo’s Human Genome Center
University of Utah
University of Wales Aberystwyth
US Geological Survey
WETA Digital*
Weill Medical College (USA)
World Radiation Data Centre
Konrad Zuse Institut (Germany)

* APJ customers

About SGI
SGI is a global leader in large-scale clustered computing, high performance storage, HPC and data center enablement and services. SGI is focused on helping customers solve their most demanding business and technology challenges. Visit www.sgi.com for more information.

Corporate Office
46600 Landing Parkway
Fremont, CA 94538
tel 510.933.8300
fax 408.321.0293
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Asia Pacific +61 2.9448.1463