Architecture, Implementation, and Deployment of a High Performance, High Capacity Resilient Mass Storage Server (RMSS)

Beata Sarnowska, Terry Jones
with
Frank Lovato, David Magee, John Kothe
NAVOCEANO Major Shared Resource Center (MSRC)
Bldg 1001, Rm 101, Stennis Space Center, MS 39529
jonestl@navo.hpc.mil
sarnowsk@navo.hpc.mil
tel +1-228/688-5344
fax +1-228/689-0400

Abstract

The Naval Oceanographic Office (NAVOCEANO) High Performance Computing (HPC) Major Shared Resource Center (MSRC) recently reengineered the existing mass storage system serving its high-performance compute platforms. The purpose was to provide significantly improved file system availability, to refresh of the technology and the architectural design, and to position the MSRC to incorporate emerging technologies such as Storage Area Networks (SANs). The resultant configuration utilizes two SUN Enterprise 10000s (E10K) with 3 TB of switched Fibre Channel Disk Arrays and the latest generation of tape devices. The theoretical peak system capacity is in excess of 3 Terabytes (TB) per day, with management of up to 1 PetaBytes (PB) of storage and an aggregate external network throughput of 220 Megabytes (MB) per second. This paper discusses the technologies and considerations used in the design of the MSRC Resilient Mass Storage Server (RMSS), its architecture, implementation and integration, deployment and transition from the existing CRAY Data Migration Facility (DMF).

1 Introduction

Mass storage servers for high performance computational centers require near-continuous availability to support applications and center throughput. The design of an HPC mass storage system must focus on resiliency as much as capacity and performance. Achieving increased data availability requires two approaches be applied simultaneously. The first is a machine-centric method titled fault-tolerance, and the second is an application-centric method called high-availability (HA).

High-Availability (HA) technology focuses on software availability through the system's ability to obtain necessary hardware resources from a pool of devices. This pool of devices is typically comprised of multiple computer systems (nodes) that form an HA Cluster. As in fault-tolerant designs, HA technology reduces Mean Time to Failure (MTTF), but also provides for designs that maintain application availability (i.e., hierarchical storage management) while portions of the cluster are removed either for maintenance and upgrade, or by failure. The disadvantage of HA Clusters is cost and complexity. The design, implementation, and operation of an HA system is more complex than a simple fault-tolerant system.

In late 1998, MSRC initiated its effort to re-engineer its existing mass storage system. Projections showed that existing technology would soon be become unable to meet user stor-

age requirements. This deployed technology was nearing the end of its useful life and support concerns left its future in doubt. This places the NAVO MSRC at risk of having no support for existing store of data, and unable to reliably support new requirements. Re-engineering refreshed the technology as well as the architectural design, incorporating fault-tolerance and HA technologies to provide a significant improvement in user filesystem availability as well as positioning the MSRC to adapt to future technologies as they become available.

For Cray sites, this paper describes a means of replacing UNICOS DMF (or IRIX DMF) with a separately hosted, resilient, highly available file server system. While it is hosted solely on SUN equipment, the Cray SV2 is scheduled to be integrated with a SunFire 6800 running Sailors 8 to support data, I/O and perhiperals, including Storage Area Network (SAN) capability. This front-end processor for the SV2 is the only place where hierarchical storage management (HSM) can be performed for the SV2. Thus, the solution presented in this paper provides a logical means to the data management functions currently performed by DMF on SV1 and earlier systems. As CRAY will not continue DMF into the SV2 O/S, a third-party solution must be found. The system implementation described in this paper provides one such solution.

2 Architecture and Design

The re-engineered RMSS Cluster is configured as a HA Cluster designed as an Active-Active, Load-Balancing Pseudo Cluster with two nodes. The RMSS cluster provides mass-storage management as the single service to client systems with each node supporting approximately 50% of the total workload. Both nodes of the RMSS cluster are comprised of E10K mainframes, symmetrically configured. The cluster management software used to create the HA cluster is Veritas Cluster Server (VCS).

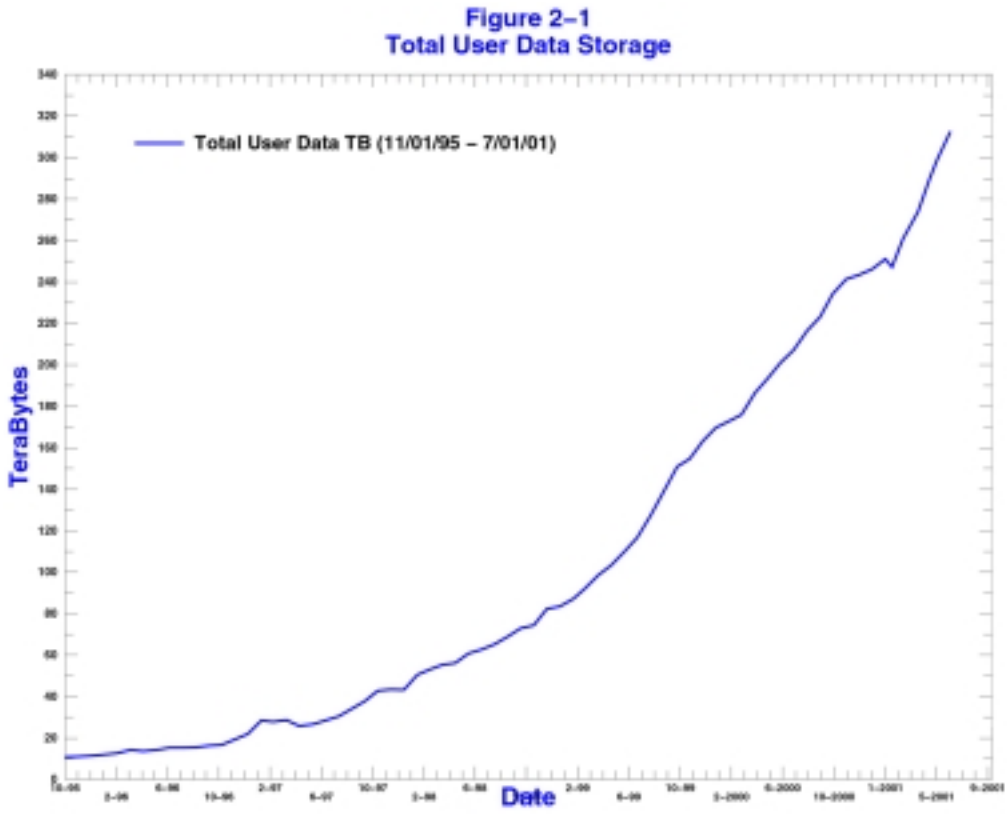
The RMSS was designed through a process which involves requirements analysis, market and product research, and engineering prototyping with loopbacks to incorporate revisions to technology and requirements. When completed, this process resulted in a full hardware and software design and specification that was then implemented using a phased approach. Each phase was structured to contain risk while incorporating a new level of capability into the system until the system functioned as required.

2.1 Mass Storage Workload Profile and Projections

The MSRC mass storage utilization and workload trends was analyzed as the first step in developing the design of the re-engineered mass storage server. Review of the historical trends showed that from December 1995 to the beginning of January 2000, the mass storage utilization typically doubled each year. The amount of data managed during the year 2000 changed with the beginning of transition to the RMSS system in October 2000. Table 2-1, MSRC Mass Storage Utilization Growth illustrates this trend.

Date	Total TB	Growth
Dec. 95	10.2	1x
Dec. 96	20.1	2x
Dec. 97	39.7	4x
Dec. 98	78.4	8x
Dec. 99	153.4	15x
Dec. 00	223.9	22x

Table 2-1
MSRC Mass Storage Utilization Growth



At the start of the transition from the UNICOS/DMF server to the RMSS, the total mass storage utilization was 225 TB of data. Currently, with 50% of the data transitioned to the RMSS, the total mass storage load is 314 TB contained in 11.6 million files. The distribution of this data varies from filesystem to filesystem; however, it follows the typical trend for most HPC centers where the majority of the data is concentrated in a relatively small number of files. At the NAVO MSRC, files larger than 500 MB account for less than 1% of all files, yet these files store nearly 70% of the total data load.

Figure 2-1, Total User Data Storage, illustrates the total data storage requirement growth since October 1995. Up until the beginning of transition of production workload to the RMSS beginning in October 2000, the storage requirement growth was very closely mod-

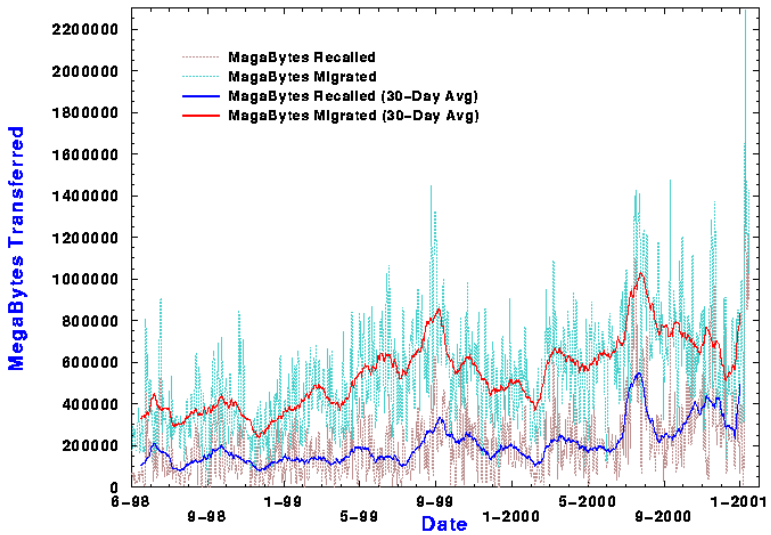
eled by a 10^{th}-order polynomial. This model considered only historical usage, but it projected the total mass storage utilization to grow by a factor of six over the initial utilization by mid-August, 2002. This alarming rate of growth was forecast to outstrip the existing mass storage and archive system's (MSAS-1) ability to support the MSRC's mission long before this point was reached.

2.1.1 Transaction Analysis

Mass storage servers are essentially transaction processors. The existing mass storage server (Cray J90/DMF based) usage history was analyzed to develop the total number of transactions and data traffic handled each since June 1998.

Figure 2-3, DMF Data Traffic, shows the number of Megabytes archived (migrated) and the number of Megabytes recalled from the archive. The system was in a steady-state until a media conversion project began in August 2000, followed by the beginning of transition of production operations to the RMSS in October 2000. The average daily transactions increased steadily until at the end of the steady-state period operations in the center required nearly 40,000 file transactions per day (1.5 TB of data); nearly all of them archive requests, which consume the most system resources in a mass storage server.

Figure 2–3
DMF Data Traffic



2.2 System Design and Analysis

Planning for the RMSS Cluster resulted in the parameters listed in Table 2-2, RMSS Cluster Design Parameters. The peak theoretical performance of ATM-622 is 622 Mbits/sec., or 77.75 MB/sec. Observed transfer rates at MSRC have been measured at 80% of the peak bandwidth, or approximately 62 MB/sec. The peak theoretical performance of serial HIPPI is approximately 800 Mbits/sec. or 100 MB/sec. Transfer rates measured at MSRC are approximately 50% of the theoretical peak or 50 MB/sec. Actual operation is anticipated to yield 60 MB/sec. transfer rates for both networks, for a total of 120 MB/sec. data

bandwidth per node. The time required to transfer 1 TB of data across the network using the actual anticipated performance of 120 MB/sec. total data bandwidth is 2.3 hours. This amounts to a 9.6% utilization of the network adapters across a 24-hour period.

Criteria	Design Node Limits	Design Cluster Limits
GB Network Traffic per Day	1024	2048
Data Archive/Recall Ratio	33%	33%
Target Disk Cache Retention Period	72 Hours	72 Hours
Largest File	25 GB	25 GB
Number of Files	25 Million	50 Million
Filesystems	2	4
Scalability (3 Year Limit)	6X	6X
Sustained Network Bandwidth	110 MB/sec	220 MB/sec
Sustained Tape Bandwidth	96 MB/Sec	192 MB/sec
Peak Disk Bandwidth	110 MB/sec	110 MB/sec
I/O Memory Bandwidth	25.6 GB/sec	25.6 GB/sec
GB Memory	16	32
CPUs	12	24
System Boards	8	16

Table 2-2
RMSS Cluster Design Parameters

2.3 Implementation

The RMSS cluster is comprised of two identically configured E10K systems (each is a node in the cluster) and three SUN T3 expansion chassis. Two expansion chassis house eight SUN T3 disk arrays each and comprise the disk cache for the SUN's Storage and Archive Manager Quick File System (SAM-QFS) hierarchical storage management software system. The third expansion chassis contains four SUN T3 series disks for metadata (inodes). Each T3 Controller Unit (CU) disk drive is connected to one port on an Ancor 8-port SANbox Fibre Channel switch. Each of the six E10K Fibre Channel adapters is connected to the other side of the ANCOR switch to allow the disk to be switched between either of the E10K systems. When VCS detects a resource failure in one of the E10K nodes, control is transferred of the disk devices, and the user filesystem on them, to the other E10K. Table 2-3, RMSS Cluster Resources, lists the set of resources available on each node and the entire cluster.

Resource	Node 0	Node 1	Total
System Boards	8	8	16
CPUs	8	8	16
Memory	16 GB	16 GB	32 GB
FC-AL adapters	6	6	12
SCSI Adapter Slots	9	9	18
System Disk	10	10	20
Tape	6	6	12
HIPPI Network Adapters	2	2	4
ATM-OC12c Network Adapters	1	1	2
Fast Ethernet Adapters	1	1	2

Table 2-3
RMSS Cluster Resources

2.3.1 RMSS Node Description

Each node is comprised of a SUN E10K server. The E10K server is a SPARC/Solaris symmetrical multiprocessor (SMP) computer system. It supports alternate pathing, providing dual paths to disk drives and network interfaces removing most single points of failure within the hardware. Hot swapable elements permit continued operations during the replacement of failed components.

The E10K can support up to 3.2 GB/sec. aggregate I/O bus bandwidth. Individual buses perform 64-bit transfers, yielding a 100 MB/sec. transfer rate per bus. However, the total centerplane bandwidth places an upper limit to this rate.

Each node of the RMSS Cluster is configured with two SUN D1000 Disk Subsystems, each with four independent disk drives, for O/S required filesystems, third-party software, and system log file space. In addition, each node is also configured with ten SUN T3 Fibre-Channel disk arrays. Eight of the T3 disks provide disk cache storage for the data in the archive filesystems and the remaining two T3 disks provide space for the metadata (inodes) for the archive filesystems. The total disk storage capacity is 1,144 GB; 144 GB provided by the D1000 disks, 1,000 GB is provided by the data T3 disk arrays and 144 GB provided by the metadata T3 disk arrays.

2.3.2 STK 9840 Tape Drives

STK 9840 SCSI-3 tape drives are installed on each RMSS node and are mounted in the STK 9411 Nearline storage silos. When RMSS was designed and engineered, the technology for tape drive failover was just reaching maturity. As a risk reduction measure, tape drive failover was not introduced into this design.

On the RMSS, typical performance of these drives under actual conditions with actual user files shows transfer rates under a SAM-QFS filesystem of 7.5 to 8.8 MB/sec. Data

compression and system workload affect the achieved transfer rates; however, the aggregate bandwidth increment for additional drives has been linear up to six drives per node.

2.4 System Software Configuration

The software for each node is identical. The operating system is SUN SOLARIS 7 and Veritas Cluster Server (VCS) is used to control the cluster. Each E10K is configured as a single domain. The central application of the RMSS Cluster is Storage and Archive Manager Quick File System (SAM-QFS) hierarchical storage management and fast filesystem package. This package is resident on each node. Also, employed on each node is Veritas Volume Manager, used primarily for its dynamic multi-pathing feature. This feature provides alternate pathing for the SUN T3 disks in the event a path failure between the node and the disk controller. Software and firmware versions are critical to the functionality and performance of the RMSS. Changing a configuration item as seemingly insignificant as the T3 controller microcode caused a complete failure of the entire cluster.

2.4.1 SAM-QFS

SAM-QFS, developed by LSC, Inc. (acquired by SUN, Feb. 2001) provides file management, storage, archive, and retrieval services under Solaris platforms. SAM-QFS provides the following capabilities:

Archive: SAM-QFS automatically copies files from the disk cache of the filesystem to magnetic tape media. Copy is performed as soon as feasible after the file has been created or modified in cache. When writing to tape, SAM-QFS combines number of small files into larger *tar* file. This allows drives to operate at streaming speeds.

Release: SAM-QFS automatically maintains the disk cache by releasing files when site-definable thresholds are reached.

Stage: When an offline (released) file is accessed, SAM-QFS will automatically copy the file from tape media back to disk cache for user access. For a sequential read of an offline file, the read tracks along directly behind the staging operation, allowing the file to be immediately available to an application without waiting for the entire file to be staged to disk cache. A stage occurs for released files when either an explicit stage command is entered or indirectly when file is referenced in a local command, or by an FTP/RCP command on a remote system. Indirect references at remote system for NFS exported SAM-QFS filesystems also causes automatic stage. For a sequential reading of multiple files staged from the same tape, files requests are sorted by tape address before recalling first file. Tape is read one time for all staged (recalled) files from the same tape VSN.

Recycle: Recycler reclaims tape space as archive copies become obsolete when replaced by newer archive versions.

SAM-QFS filesystem is implemented using the standard SUN/Solaris virtual filesystem (vfs/vnode) interface. By using the vfs interface, QFS works with the standard Solaris kernel and requires no modifications within the kernel for file management support. It must be tested and verified with each Solaris release.

SAM-QFS supports multiple QFS filesystems with up to 200 partitions each. Disk space is allocated in fully adjustable Disk Allocation Units (DAUs) from 16 to 65,535 1K-byte blocks. The bandwidth of a QFS filesystem has been measured at over 1 GB/sec. Its peak theoretical bandwidth is stated to be 1.5 GB/sec. SAM-QFS supports striping (RAID-0) or round robin disk allocation. Round-robin allocation is used on the RMSS QFS filesystems.

The number of files contained within a QFS filesystem is limited by the amount of disk storage space available for inodes (metadata). Inodes are dynamically allocated and are 512 bytes each. QFS stores the inodes on the metadata device that is a separate device (for increased performance) from the file data devices. QFS is a 64-bit filesystem and files may be up to 2^{64}-1 bytes. SAM-QFS data is written to archive tapes using the standard tar format. It allows data to be read and recovered from tape archive outside SAM-QFS environment using *gnutar* utility.

2.4.2 SAM-QFS Filesystem Configuration

The QFS filesystems store the entire MSRC user file base. The disk resources allocated to the QFS filesystems hold the online portion of the filesystems and act as a cache mechanism governed by the Least Recently Used algorithm. The metadata (inodes) for offline files are always completely online, on a separate disk array unit. Each file typically has one metadata entry (an exception is symbolic link for which a file will have its own inode plus as many inodes are there are links to it), and each metadata entry is 512 bytes in size. Metadata entries contain information about the files, including location, size, access date and time, and archive and recall attributes. Integrity of the metadata is critical to integrity of the data. If the metadata is corrupted or lost, it is possible to lose all files managed by the Cluster. Integration testing concentrated heavily on developing highly reliable routines and procedures to back up the metadata and to frequently check its integrity.

The RMSS Disk Subsystem is comprised of SUN T3 disk arrays, a high-performance, modular, scalable, RAID storage device containing nine internal 18.2 GB disk drives for up to 162 GB of storage capacity per tray. These disks provide the metadata and the disk cache space (online portion of the filesystems) for the RMSS Cluster. Filesystems are not shared between the nodes. Each disk unit and the filesystems contained on it are mounted on only one node at any given time.

The T3 disks are connected to E10K systems using six Ancor SANbox SL 8-port FC-AL switches for failover access, as well as providing Dynamic Multi-Pathing (DMP) support for continued access without triggering a failover to the partner host in the event of a switch or HBA failure. Veritas Volume Manager provides DMP support to the Solaris O/S.

2.4.3 Network Connectivity

The default network for external access to either node of the RMSS cluster is ATM-OC12c with a peak bandwidth of 622 Mb/s (77.75 MB/sec.). Each node is also configured with two serial-HIPPI network interfaces at 100 MB/sec. each. One interface is connected to the internal MSRC network for bulk data transfers between the RMSS and its client HPC systems. The other interface is dedicated to the transition of data from the existing DMF mass storage server to the RMSS.

Each node is also configured with multiple fast ethernet ports for internal connectivity to support RMSS functionality. Primarily, these fast Ethernet are used for System Support Processor (SSP) to E10K, and E10K to High Availability Automated Cartridge System Library Software (HA-ACSLS) communications.

2.4.4 STK High Availability Automated Cartridge System Library Software (HA-ACSLS)

The STK HA-ACSLS platform used in RMSS is a packaged configuration from STK, Inc. which uses two SUN Ultra-10 workstations to create an HA environment for the ACSLS tape library management software. The HA-ACSLS cluster is designed as an Active-Passive Cluster of two nodes (Pseudo-Cluster). Further, the cluster includes two STK 9330 Library Management Units (LMUs), each with dual path connections to each Ultra-10 workstation. This provides a complete high-availability environment for all STK robotic tape library management and control.

The tape library database for the HA-ACSLS software resides on a STK 9133 (27 GB) RAID disk array with dual internal storage processors (controllers) configured for resiliency. The disk array is manufactured by Data General, Inc. under the name Clariion and sold by STK. Internal failover between the dual storage processors is handled by the Clarion Application Transparent Failover (ATF) software. Ultra-10, LMU and Ultra-10 peripheral and network adapter failover is accomplished by Veritas Cluster Server (VCS).

SAM-QFS and HA-ACSLS inter-operate with each other in a client-server relationship to form the complete mass storage system. HA-ACSLS acts as the server for the Automated Cartridge System (ACS) Library. It presents the control interface to clients to request tape mounts and dismounts of the robotics only. Data flow is not managed by HA-ACSLS.

The HA-ACSLS client software resides on the E10K nodes requesting tape mounts and status of the HA-ACSLS. It is important to note that SAM-QFS does not communicate with the HA-ASCLS for anything other than tape mount/dismount requests. It treats the HA-ACSLS strictly as a black-box server and the only status information exchanged is the completion code of the command it issued. SAM-QFS keeps its own catalog of tapes and does not query HA-ACSLS for status or state information. This had special implications on the design of the VCS failover scripts, which must reestablish machine state.

2.4.5 Veritas Cluster Server (VCS)

VCS is the software package that provides the resiliency of RMSS Cluster. VCS detects service-level failures on either node in the cluster, notifies operators and system administrators, and institutes recovery procedures to shift the service from the failed elements of the cluster to reserve functioning elements [6]. VCS is employed on both the RMSS Cluster and the HA-ACSLS Tape Library Subsystem.

VCS uses agent scripts and programs as the intermediary between a service and VCS. Custom-developed agent scripts were required to accomplish the failover functionality of SAM-QFS. These scripts were developed during the configuration of VCS. Discussions held with the Aeronautical Systems Command (ASC) MSRC regarding these scripts identified several important considerations with regard to SAM-QFS state at the time of failover.

3 Performance and Availability

During unit and integration testing, a number of functional and performance tests were conducted. These tests assured the functional correctness of each stage of the integration as well as provided performance measurements to assure that the RMSS would support the workload targets it was designed to handle. This section presents results from most significant of these performance results.

3.1 SAM-QFS Single File Functional Tests

These tests measured the time required to write a large, incompressible file from disk cache (T3) to tape (9840) in an archive operation and to recall that file back to disk cache from offline status in a stage operation. The results are listed in Table 3-3, SAM-QFS Functional Tests Results.

Operation	File Name	File Size (Bytes)	Data Transfer (sec)	Tape Position (sec)	Total Time (sec)	Bandwidth (MB/Sec)
Archive	5gb_binary	5,368,971,264	3,334	8	3,342	1.53
Archive Next	5gb_binary	5,368,971,264	3,436	0	3,436	1.49
Stage (Recall)	5gb_binary	5,368,971,264	602	10	612	8.51

Table 3-3
SAM-QFS Functional Tests Results

The first time the file is copied from disk cache to tape, the archive tape was already mounted, but not positioned to the write location. The positioning took 8 seconds, typical for the STK 9840 tape drive. After the file was written to the tape, it was deleted and a "new" version of the file was created and archived. There were no measurable performance differences for the transfer except for the lack of tape positioning time when the second file was archived. The performance of 1.5 MB/sec is worst case for the STK 9840 and results from writing a file that is completely uncompressible and is too large (greater than 8MB) to fit in the tape drive's buffer. The effective bandwidth on recalls back to disk cache (stages) approaches the typical data transfer rate for data from STK 9840 tape drives to a SAM-QFS filesystem mounted on unstriped T3 Disk Arrays. The SUN E10K, Solaris 7 and unstriped T3 disk drive are capable of considerably higher data transfer rates; however using the STK 9840 tape drive and a 256K tape block size limits data transfer rates to approximately 8.5 to 8.75 MB/sec. This bandwidth is consistent with average performance seen on the RMSS under production workloads.

3.2 Profiled Data Set Performance Tests

These tests were performed to determine the total RMSS system response time to a workload representative of actual conditions. The existing filesystem file size distribution on the original Mass Storage (MSAS1) was profiled to determine the counts of files within given filesize ranges. A test data set was then constructed patterned after the distribution percentages for each range applied to a fixed total number of files. Actual user files were

then extracted from the MSAS1 server to create this composite set with the same proportion of files in each range to create a group of files that represented the actual file distribution of the production workload. All of these files together created a test set of nearly 6 GB in size.

Tests conducted with the profiled data set involved single and multiple stream transfers across direct ATM OC12c network. The source system was RMSS node. The test results in this section were for a multiple stream test which was designed to closely simulate actual workload conditions for the RMSS Cluster. Table 3-7, Aggregate Stream Performance, shows the composite time for all four streams combined.

Operation	Elapsed Time (Sec.)	Bandwidth (MB/Sec.)	Average (Sec./File)
Transfer Data	407	14.93	0.98
Archive Data	670	9.07	1.61
End-to-End	776	7.83	1.86

Table 3-7
Aggregate Stream Performance

The results show that, the RMSS could easily sustain at least 28 GB/Hour for a total of 672 GB/Day using the ATM-with only two tape drives per node. The production configuration has seven STK 9840 tape drives for Node 0 and eight STK 9840 tape drives for Node 1, with the ability to increase these by a factor of three and an STK 9940 tape devices. For the same amount of profiled data, using the same overlap between network to disk transfer time and disk to tape transfer time (72%), the production configuration can be shown to accommodate the stream in 216 seconds, a bandwidth increase of 3.6 times.

Applying this derived scaling factor of 3.6 the end-to-end bandwidth of 28 GB/Hour attained in the test, the RMSS Cluster is projected to sustain a peak of at least 100 GB/Hour or 2.4 TB/Day. This is sufficient to meet all requirements and the design criteria for a cluster capability of 2 TB of traffic/day as listed in Table 2-2, RMSS Cluster Design Limits.

In production, sustained periods of 50 MB/sec transfer rates across ATM are typical. This is in excess of 3 times the measured test dataset transfer rate of nearly 15 MB/sec. Analysis of the production system indicates that the ATM network performance is being limited from reaching the design specification of 62 MB/sec per node by the number of TCP/IP send and receive buffers available. A planned increase in buffers is anticipated to increase the maximum attainable network bandwidth to the design specification.

4 Transition to Production and Future Work

Moving over 225 TB of data from one HSM format (DMF) to another (SAM-QFS) without interruption to the user community required careful planning and construction of special routines. Risk management, continued availability to the user community, requires that no configuration changes be made, including upgrading of software and hardware components during the transition. The only exception is the application of critical patches

and microcode to correct experienced hard errors. Future work must therefore wait until the data has been transitioned.

4.1 Transition to Production

The ideal archive transition approach, would be to step-wise un-cable the *data-pipes* from the old archive host server and re-cable them to the new host server(s). The data archive would be common, and would be continuously accessible to users as they transitioned to use of the new archive server.

For NAVO MSRC, this approach would have at its core the conversion of Cray DMF *encapsulated* data, databases and media encode, into something understood by the SAM software managing the new server hierarchical storage and filesystem architecture. Implementation details make this *ideal, re-cable* approach utopian and not very practical.

Various other alternatives for data archive migration were considered. These included recalling and converting over 225 TB in a single dedicated period of time, moving data User by User, moving data Group by Group, or other logical groupings such as by VSN, by project, etc. All were determined to be technically impractical and administratively complex on a scale as large as 225 TB. After considering all of these alternatives, it became obvious that the MSRC must avoid subjecting users to any uncertainty about the location of their data (MSAS1 or RMSS node) and any significant periods (4 to 8 hours or more) where their data is unavailable.

The method chosen is a user-centric approach, where data is moved to the RMSS node (its new location) from MSAS1 when it was needed. This led to a review of LSC's Migration Toolkit, (MigKit). As delivered, MigKit had examples, which supported the establishment of local SAM files whose data are resident on foreign media, not SAM controlled local archive media. The specific example, for which source code was provided, is for retrieval of a file from a CD in a locally accessible CD ROM. A simple modification of this example was made to change the reading a file from a CDROM, to reading a file using a simple *rsh* across a network.

Prior to the start of the transition of a given filesystem, the MSAS1 server is taken into dedicated mode and the directory structure for the filesystem to be transitioned is replicated onto the RMSS system. The MSAS1 system is then returned to regular access with the exception of those files were located in the affected filesystem. Access to these files is now managed through the RMSS. All of the directory entries on the RMSS are initially created as *foreign* files. Once this process is accomplished, users and applications will use the RMSS as the file server, no longer communicating with the old MSAS1 system. Whenever an application accesses a file, the SAM foreign file mechanism, supported by rsh coding, retrieves the file contents from MSAS1 and writes it onto the RMSS. The file is delivered to the user application and, as part of the transfer, the file is flagged for re-archive as a native SAM file on local SAM controlled media. It is then permanently located on the RMSS node. The next request to read or modify the file will take place completely within the RMSS node. In this manner, access to users files is maintained throughout the data transition process, without the users being required to know the status of their files.

Several important attributes are inherent to this approach. Any file on the new server typed as a foreign file has not been moved from the old archive and is a candidate for migration. Conversely, any file that is a native file type has been migrated to the new server and is no longer needed on the old archive. This greatly simplifies the accounting and progress reporting for the transition. Additionally, this also simplified the strategy for implementing a bulk data transfer capability on behalf of users. In the background, an optimized automated file *staging* routine runs by staging packets of foreign-media-resident files all residing on a common media. This process works by mounting a DMF tape one time, then transitioning all files on that DMF tape to the RMSS.

Transition began with files owned by internal users (MSRC staff members) in July of 2000. Over 4.5 TB of data were moved during a low-intensity process lasting approximately two months. During this data movement, the transfer and bulk move routines were "modified" and proven. In mid-October, 2000, transition began for the first large group of MSRC users. This phase lasted 2 months during which over 20 TB, nearly 92 percent of the total data volume was moved. The remaining eight percent was low-priority files that may be candidates for deletion. The transfer rate achieved an average of 660 GB/day.

In mid-January, 2001, the second large group of MSRC users began the transition process. This filesystem supports the majority of the user community. This work was completed in approximately four months. Improvements in the process developed during the transition of these files resulted in an improved transfer rate of 1 TB/day.

The final filesystem to be transitioned consists of the majority of the data to be moved. This filesystem contains over 60% of the total data stored in the MSAS1 system (154 TB). The starting date for the transition of this filesystem has not been set as of the writing of this paper. A desired goal to complete the transition from the MSAS1 server is September 30, 2001. Support overhead for administration and maintenance of the migration is 1.5 analysts. Transition is pending receipt of additional tape resources, but is anticipated to began during 4Q FY2001.

5 Conclusion

The need to maintain a stable configuration is balanced with the requirement to install new software levels. As the RMSS is becoming the production mass storage server for the MSRC, integration of new technology will occur cautiously so as to not disrupt center operations or risk the integrity of the data. Anticipated future upgrades include higher capacity data cache fibre channel tape devices, tape devices optimized for bulk storage of large files and a simplification of the fibre channel topology and system software.

The MSRC RMSS has proven itself capable of handling the workloads it was designed to handle and is flexible enough to scale to meet its projected requirements. The success of the RMSS is also measured by the fact that the all four MSRCs are implementing versions of this design. Discussions are also being held to develop a smaller scale version for other DoD HPC centers and interest has been expressed on the part of non-DoD government entities.

Acknowledgements

Many were involved who contributed to the success of this project. In particular, the ASC MSRC began implementation of the RMSS Cluster concept early and uncovered many issues. Through a sharing of information, the MSRC was able to avoid these problems. Much gratitude is extended to the ASC MSRC for their efforts.

References

1) *Application Storage Manager (ASM) System Administrator Guide*, Release 3.3.1, Storage Technology Corporation, Louisville, CO, (C) June 1999.

2) *Sun StorEdge T300 Installation, Operation, and Service Manual*, Sun Microsystems, Inc., Palo Alto, CA, (c) October 1999, Part No. 806-1062-10, Revision A.

3) http://www.ancor.com/prod.html

4) *9840 Tape Drive*, Storage Technology Corporation, Louisville, CO, (c) 1999, http://www.stortek.com/StorageTek/hardware/tape/9840.

5) largefile(5) Solaris Man Page, (c) Sun Microsystems, Inc.

6) *Veritas Volume Manager for Solaris, Getting Started Guide*, Release 3.0.1, Veritas Software Corporation, Mountain View, CA, (C) May, 1999, P/N 100-001123