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

_

Terry Jones, Beata Sarnowska
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 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 2 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. Existing mass storage server designs are capable of handling near-petabytes of data but are not resilient and therefore, become a single point of failure for the entire complex of HPC systems. Additionally, implementation of system enhancements, corrective hardware and software, routine maintenance and system failures result in outages which affects the entire complex of HPC systems. Because of the high level of investment in these systems and the critical nature of the applications executing on them, any such outages are extremely expensive. To mitigate these impacts, 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.

Fault-tolerant computer systems have been available since the early 1980s, and focused heavily on both improving the reliability of the system's components as well as providing for hot-spares, which can be brought into service automatically in the event of a primary component failure. Fault-tolerance has been chiefly employed to improve the availability of online transaction processing and interactive database applications. Fault-tolerance reduces the Mean Time to Failure (MTTF) for a system; however, it cannot provide resiliency for non-disruptive hardware upgrades and maintenance.

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 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. In some applications, two servers are installed with one often acting as a "hot-spare". The design, implementation, and operation of an HA system is more complex than a simple fault-tolerant system.

At the lower end of the mass storage server performance spectrum, mass storage servers have achieved higher levels of availability, but HPC centers require high-end performance mass storage servers capable of both high capacity (up to petabytes and a hundred-thousand data requests per day) and high external bandwidth (200+MB/sec sustained) and high data volume (2+ TB/day). Functional availability of 99.99% for HPC mass storage servers has yet to be achieved but combining existing and new technology has made it possible to design HPC mass storage servers that closely approach these levels of availability.

With careful design and selection of technology and implementations, such a design can be implemented and positioned to take advantage of emerging technologies, such as robust implementation of Fibre Channel and full-function SANs, high-speed and high-capacity tape devices, fibre channel tapes and large capacity disks for filesystem caching. In late 1998, MSRC began the effort to re-engineer its existing mass storage system. Projections showed that the current technology would soon be outstripped by user storage requirements, requiring near-heroic measures to continue efficient center operation. Further, the deployed technology was nearing the end of its useful life and uncertain vendor support left its future in doubt. The center was at risk of having little to no support for its existing store of data, and unable to reliably support new requirements. The 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.

2 Architecture and Design

The RMSS was designed through a process involving requirements analysis, market and product research, 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 incorporate a new level of capability into the system until the system functioned as required while at the same time minimizing risk. To contain the risks inherent in such a new and complex design, each phase had specific goals that were required to be achieved before work could begin on the next step.

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).

2.1 Mass Storage Workload Profile and Projections

An analysis of the MSRC mass storage utilization and workload trends was undertaken as the first step in developing the design of the re-engineered mass storage server.

2.1.1 Mass Storage Utilization

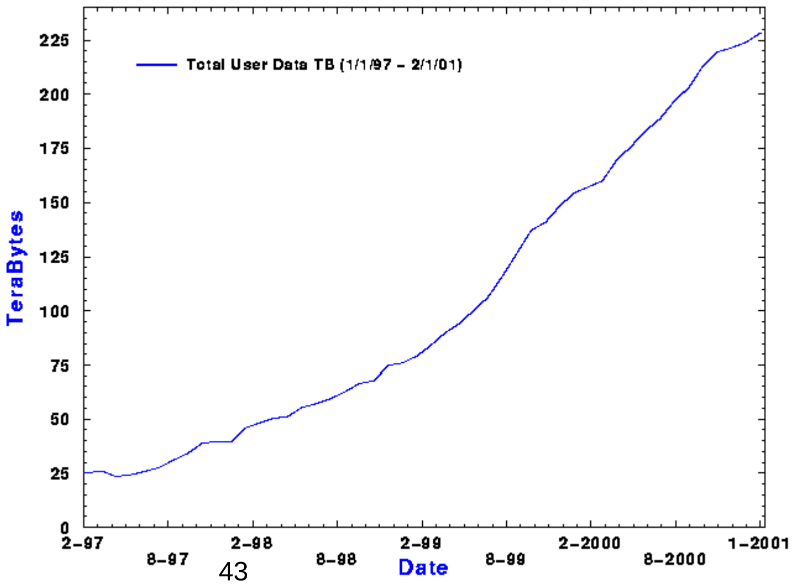
A review of the historical trends shows that from December 1995 to the beginning of January 2000, the mass storage utilization typically doubled each year. The amount of data managed during the year 2000 changed with the beginning of transition to the RMSS system in October 2000. Table 2-1, MSRC Mass 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 transition from the DMF based server to the RMSS, the total mass storage utilization was 221 TB of user data. The distribution pattern for this data roughly followed a 90/10 rule where 90% of the data is contained in 10% of the files. A single project

Figure 2-1
Total User Data Storage



accounted for nearly two-thirds (60%) of all data managed.

2.1.2 Workload Projections

Figure 2-1, Total User Data Storage, illustrates the total data storage requirement growth since February 1997. Up until the beginning of transition of production workload to the RMSS beginning in October 2000, the storage requirement growth was very closely modeled by a 10th-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.

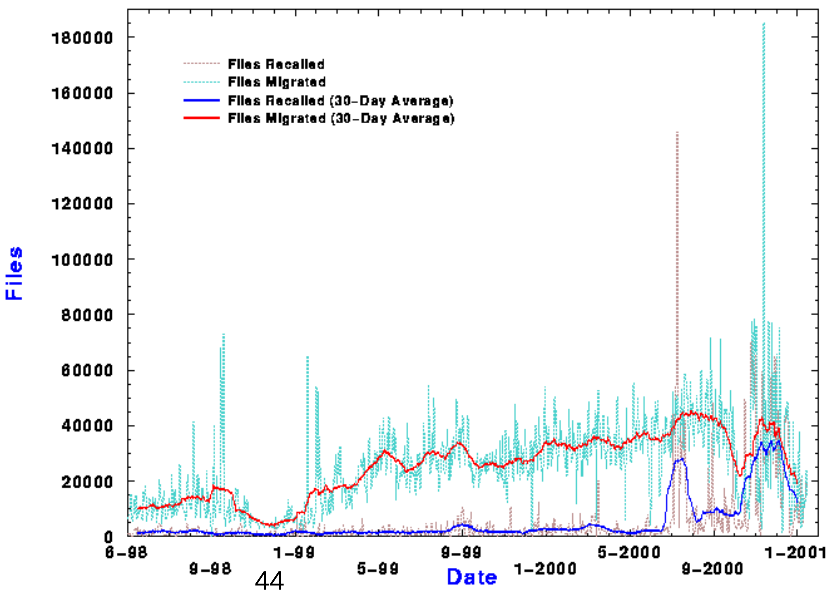
Additional increases in data storage requirements are anticipated as the accuracy and resolution of several key applications is increased. Two-dimensional simulations increase the storage requirements by the square of the resolution increase while three-dimensional simulations present a cubic relationship between storage requirement and resolution increase. Anticipated key application improvements were factored into the analysis, yielding a projected six-fold increase over initial utilization by March 2002.

Should additional computing resources be added to the MSRC, a six times increase may occur earlier than projected. However, there is evidence to suggest that Moore's Law of computational power increases are implicitly part of the storage model. If this is the case, any growth in computational resources has already been considered by the storage model. This lends credence to the design goal that calls for the capacity of the RMSS to be able to be expended by a factor of six over its lifetime.

2.1.3 Transaction Analysis

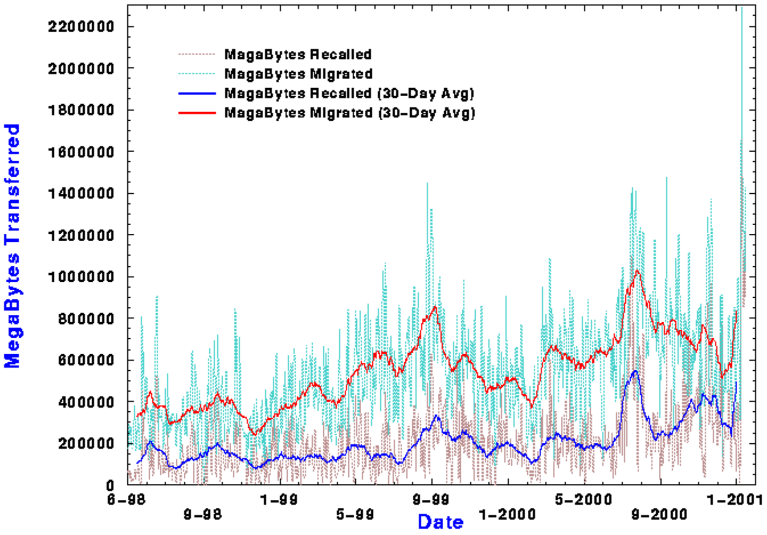
Mass storage servers are essentially transaction processors. The transactions they handle are of varying duration, depending on the location of the file (in disk cache or offline on tape) and the size of the file. Each transaction requires either an amount of data (a file) to be accepted for storage (a put or an archive request) or delivered to a requesting application (a get or a stage request). 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 2-2, DMF File Traffic, shows the number of files archived (migrated)

Figure 2-2
DMF File Traffic



and the number of files 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. A reduction in transitions occurred at the end of 1998 and the beginning of 1999 when the disk cache on MSAS1 was increased to improve system response time. 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; nearly all of them archive requests, which consume the most 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. These criteria were developed with a review of workload characteristics, input from management, user support, user allocations and allocation plans, Original Equipment Manufacturer (OEM) and consultant input. These criteria were used to tailor the integrated configuration to best meet the MSRC's immediate and longer-term requirements.

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 (02:18:53). This amounts to a 9.6% utilization of the network adapters across a 24-hour period.

The default blocksize that the Quick File System (QFS) software uses for STK 9840 tape devices is the maximum blocksize of 256 KB. This rate provides optimum streaming performance.

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	5	5
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	60 MB/sec	60 MB/sec
I/O Memory Bandwidth	25.6 GB/sec	25.6 GB/sec
GB Memory	8	16
CPUs	8	16
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 called a node in the cluster) and three SUN T3 expansion chassis. Two expansion chassis house eight SUN T3 disk arrays each comprise the disk cache for the LSC'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. 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. Switching between E10K systems occurs in the event of VCS detected failure within one of the E10Ks. This transfers control of the disk devices, and the user filesystem on them, to the other E10K. Figure 2-4 illustrates the configuration of the RMSS cluster.

FIGURE 2-4 – RMSS CONFIGURATION

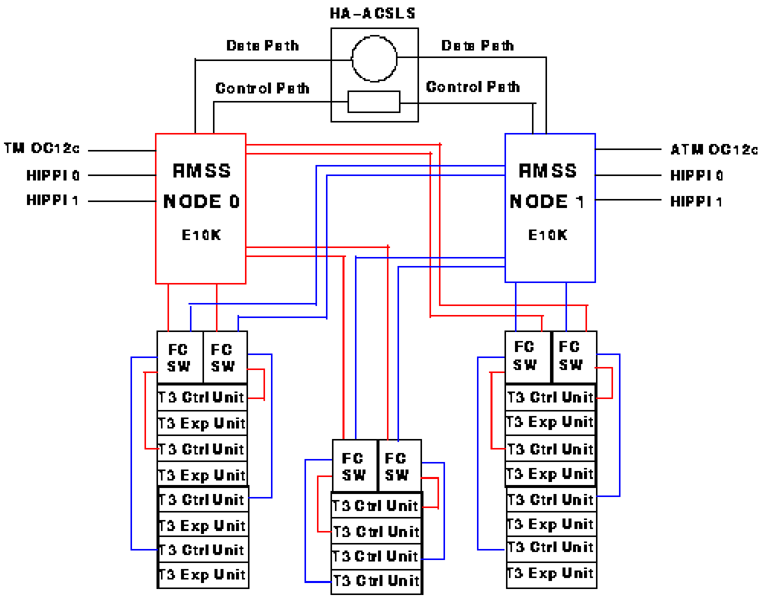


Table 2-3, RMSS Cluster Resources, lists the set of resources available on each node and the entire cluster.

Resource	Node 0	Node 1	Total
System Boards	8	8	16
CPUs	8	8	16
Memory	8 GB	8 GB	16 GB
FC-AL adapters	6	6	12
SCSI Adapter Slots	9	9	18
System Disk	3	3	6
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

The E10K system is a SPARC/Solaris symmetrical multiprocessor (SMP) computer system, designed as a general-purpose applications and data server for host-based or client/server applications such as communications and data management. The E10K utilizes alternate patching and redundant architecture to remove most of single points of failure within the hardware. This was a significant reason for its selection as the node component of the RMSS. Hot swappable elements permit continued operations during the replacement of failed components. The system supports alternate pathing that provides dual paths to disk drives and network interfaces.

The centerplane contains dual-redundant 16 x 16 crossbar switches that allow all system boards to independently and simultaneously exchange data over independent data busses with a maximum theoretical bandwidth of 25.6 GB/sec. System boards connect the CPUs, main memory, and I/O subsystems to the centerplane. Each system board can support up to 4 CPUs (400 MHz), 4 GB of main memory interleaved into four banks and two I/O buses (SBus or PCI). A system board can be configured with only one type of I/O bus. SBus configured system boards may contain up to 4 slots for SBus peripheral adapters (two per bus). PCI configured system boards may contain up to two slots for PCI peripheral adapters (one per bus). Each system board has four 100 MB/sec paths to the centerplane, each transferring 16 data bytes per bus cycle. Two paths are for CPU access, one path is for I/O data access and one is for memory access. There can be up to 16 system boards.

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.

Each D1000 disk drive provides 18 GB of storage, and each T3 disk array used for data is configured to provide 125 GB of RAID-5 storage. The T3 disks configured for metadata provide 72 GB of RAID 1+0 storage each. 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 SAM-QFS

Storage and Archive Manager Quick File System (SAM-QFS), developed by LSC, Inc., is a hierarchical storage manager and a very fast filesystem software package that 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. Copies are made after a specified time period has passed since the file was last accessed. Up to 4 copies of a file can be made. Archiving is typically performed as soon as feasible after the file has been created or modified. SAM-QFS can be directed to archive files on a filesystem, directory, or even a singe-file basis. It can also be configured to archive files immediately or to mark files to never be archived. Users can also immediately archive any file owned by their userid.

Release: SAM-QFS automatically maintains the disk cache at a specified percentage use threshold. Two threshold values are used to manage usage of the online disk cache. When online allocated disk space exceeds the high percentage usage threshold, SAM-QFS automatically begins releasing archived files from the disk cache. Disk cache space is released until the low percentage usage threshold is reached. The order in which files are considered for release from online disk cache is determined by two site-definable weighting factors for file size and file modification age.

Stage: When an offline (released) file is accessed, SAM-QFS will automatically copy the latest version of 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. The user can direct SAM-QFS to immediately stage any file that their userid owns.

Recycle: As users modify files, archive copies associated with the old versions can be purged from the media. The recycler component of SAM-QFS identifies the tape volumes with the largest proportion of expired archive copies and directs moving non-expired (most recent version) to different volumes. [1]

SAM-QFS filesystem is implemented using the standard SUN/Solaris virtual filesystem (vfs/vnode) interface. The kernel intercepts all requests for SAM-QFS resident files and passes the request to QFS. By using the vfs interface, QFS works with the standard

Solaris kernel and requires no modifications within the kernel for file management support.

SAM-QFS supports multiple QFS filesystems with up to 200 partitions each. Disk space is allocated in Disk Allocation Units (DAUs). QFS supports a fully adjustable DAU, from 16 to 65,535 1K-byte blocks. This adjustable DAU is useful for tuning the file system with the physical disk storage device, eliminating system overhead caused by read/modify/write operations [1]. The Sun T3 disk array supports 16, 32 and 64 KB DAUs, on the hardware level, for physical allocation. The choice of the T3 DAU is very important to the performance of the disk array as it impacts the way cache memory within the device is used to avoid expensive read/modify/write operations. Tuning the T3 disk array DAU to the characteristics of the files within the filesystem and then tuning QFS to the T3 disk array DAU can make a significant impact on filesystem performance. 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. Striping spreads the file allocation across the file system devices while round robin allocates files on individual file system devices. Striping is useful when the additive performance of all devices is needed for a single transfer. Round-robin allocation is used when the aggregate performance is required for high-performance multiple-stream transfers. Round-robin allocation is used on the RMSS QFS filesystems.

The number of files contained within a QFS filesystem is limited only 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 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.

2.3.3 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 most recently accessed files are likely to be found online. All other files are most likely residing on tape, in the offline portion of the filesystem. The metadata (inodes) for these filesystems are always completely online, on a separate disk array unit. Each file typically has one metadata entry (an exception is symbolic links 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 initial QFS filesystem configuration was carefully designed to balance performance with resiliency, the primary consideration being resiliency and data availability. It considers data usage patterns and file characteristics, as well as segmentation of the user data into separate filesystems to localize the effects of filesystem failures. The hardware architecture and failover requirements are also considered in this design to provide a balanced workload and rapid recovery in the event of a failover. Table 2-4, Initial QFS Disk Cache Filesystem Allocations, presents the filesystem design details.

File System	Family Set Name	Filesystem Attributes	Filesystem Attributes	Portion	Number of Inodes	Node 0 Data GB	Node 0 Devices	Node 1 Data GB	Node 1 Devices
/u/a	samfs1	Large # of Files Small File Sizes	Low # of Requests Low B/W	Metadata	140.6 M	72 GB	c7t2d1		
				Data		252 GB	c6t3d3 c6t3d1		
/u/b	samfs2	Small # of Files Large File Sizes	Low # of Requests High B/W	Metadata	140.6 M			72 GB	c7t2d1
				Data				504 GB	c3t3d3 c3t3d1 c1t1d2 c1t1d0
/u/c	samfs3	Large # of Files Mixed (90/ 10) File Sizes	High # of Requests Avg. - High B/W	Metadata	140.6 M	72GB	c3t1d0		
				Data		756GB	c5t1d2 c5t1d0 c6t3d3 c6t3d1 c4t1d2 c4t1d0		
/u/d	samfs4	Large # of Files Small File Sizes	High # of Requests Avg. B/W	Metadata	140.6 M			72 GB	c4t1d0
				Data				504 GB	c5t3d1 c5t3d3 c1t1d2 c1t1d0

Table 2-4

Initial QFS Disk Cache Filesystem Allocations

The RMSS Disk Subsystem is comprised of SUN T3 disk arrays. These disks provide the metadata and the disk cache space (online portion of the filesystems) for the RMSS Cluster. The disks are accessible by either node in the cluster through the 8-port ANCOR SANbox Fibre-Channel Arbitrated Loops (FC-AL) switches; however, 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 data disks are configured so that each cabinet houses two FC-AL switches and one T3 partner group for each RMSS node. Each partner group is made up from two sets of disk arrays comprised of a T3 Controller Unit (CU) and a T3 Expansion Units (EU) device. This provides an alternate path through the

FC-AL switch to all disks in a partner group in the event of a T3 CU controller failure. The two FC-AL switches in each cabinet provide alternate paths to mitigate the failure of a switch, Host Bus Adapter (HBA), or system board within a node and reduce the requirement to failover to the other node.

The metadata devices are all T3 CU and are all contained in the same cabinet. Two are looped together for each node to form a partner group. Two switches are also present in the cabinet to provide accessibility to both nodes as well as alternate pathing.

The T3 Disk Array, as deployed on the RMSS, is 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. Newer models contain even larger capacity disk drives for greater total storage per unit. The T3 connects to the host node via FC-AL over copper media. A Media Interface Adapter (MIA) is required to adapt the copper FC-AL media to the E10K's optical FC-AL media.

The T3 units deployed in the RMSS were pre-production units, consisting of CU and EU units. T3 CUs are configured with an internal RAID controller, providing the capability to act as a standalone device. The T3 EU is identical to T3 CU except that it does not have the internal RAID controller and must be connected to (hosted by) T3 CU. However, SUN never released T3 EU for general availability and there are no plans to do so. T3 CUs can be connected together to provide an additional connectivity path and a total of 512 MB cache memory. [2]

The T3 disks are connected to SUN (Qlogic) 6730A FC-AL HBA on both E10K systems using six Ancor SANbox SL 8-port FC-AL switches. These switches connect all T3 disks to both E10K hosts 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. Ancor switch architecture allows large heterogeneous SANs to be constructed without the performance degradation. Ancor switches will operate in this multistage, multiswitch environment as well as a cascaded environment. Low fabric latency of less than 600 nanoseconds (best case, no contention) and cut-through routing allows for high-performance routing in either architecture.

Ancor switches use a 4-port Application Specific Integrated Circuit (ASIC) with embedded memory, enabling on-chip data transfer without accessing shared memory. Shared memory technology performance becomes increasingly constrained as the number of processors accessing memory. Embedded memory eliminates the contention issue and allows for large switches and large SANs without performance degradation.

With full-fabric support, the Ancor switch supports private and public loops. The switch is protocol independent and connects with UNIX, NT, Linux, IBM, DEC and Macintosh systems. [3] The Ancor switches used in the RMSS are early access models and are deployed in Hub-Emulation mode. The switches are mounted inside the T3 disk enclosure racks. Each enclosure supports up to eight T3 disks along with two Ancor switches.

2.3.4 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. A significant feature of the 9840 is the tape media contains two internal spools; a tape spool and a take-up spool. The tape is positioned to the middle of its length before unmounting so that the tape is always at center point for the next mount. This has the advantage of reducing the average search time by 50% over single-reel cartridge media. Specifications for the STK 9840 are listed in Table 2-5, STK 9840 Specifications. [4]

Specification Item	Value
Tape Speed Read/Write	2 meters/sec
Tape Speed Search/Rewind	8 meters/sec
Tape Load/Init to Ready	4 sec
Search Time	8 sec (1st search) 11 sec (average)
Max Rewind Time	16 sec
Interfaces/ Drive	1 ESCON or 1 Ultra-SCSI (FC-AL pending)
Compression	Enhanced LZ-1, 4:1 ratio, (application dependent)
Data Buffer Size	8 MB per drive
Bandwidth, Head-to-Tape Peak Rate	10 MB/sec (application dependent)
Bandwidth, Inference Peak Rate	18 MB/sec ESCON 40 MB/sec Ultra-SCSI (application dependent)
Maximum Blocksize	256 KB

Table 2-5
STK 9840 Specifications

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.

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. Anticipated improvements in hardware technology, such as FC-AL tape drives, have materialized and this capability now exists. As deployed, the RMSS requires one tape drive per SAM-QFS filesystem, and when the filesystems are failed over to the other node, all tape drives on the failed node are lost to the cluster. The incorporation of tape failover is planned.

2.3.5 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 for the RMSS Cluster is SAM-QFS hierarchical storage management and fast filesystem package. This package is resident on each node. The configurations are slightly different on each due to the slightly asymmetrical workloads. 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.

2.3.6 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. Only three systems participate in this private, switched network: both of the two RMSS nodes and the DMF server.

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

The public names for each node are assigned to the default network ATM-OC12c addresses. Only the public names are listed in the Domain Name Servers (DNS). Both nodes have access to DNS and are able to transfer outbound traffic to any machine supporting kerberized ftp, and rcp connections.

2.3.7 Security

As the central repository for all user data at the MSRC, data integrity and security are the principle factors guiding the design and integration of the RMSS Cluster. Data stored on the RMSS often represents years of research and substantial investment on the part of the agencies using the MSRC systems. It is critical that this data be protected to the maximum extent possible, meeting and/or exceeding the existing and anticipated requirements and regulations of all relevant governmental agencies involved. The security risks, which must be addressed by a resilient mass storage cluster such as the MSRC RMSS, are:

General System Security: Patch levels, inetd configuration, Protection against unauthorized access (Kerberos), Data Transmission Encryption, Immunity from External attacks (e.g. Denial of Service Attacks)

Backend Network: SSP/E10K, E10K/HA-ACSLS communications, Isolation from Public Networks, Unauthorized Access, Immunity from External attacks (e.g. Denial of Service Attacks)

Data Security: Protection of Data Storage Tapes, user access to tapes and tape devices, protection against unauthorized access to SAM-QFS archive tapes, encryption of transmitted data where required.

All relevant DoD, HPCMO, NAVO, and other government agency regulations and requirements were implemented in the RMSS Cluster. All vendor-recommended patches and updates were installed during initial O/S configuration. Kerberos along with secure shell (ssh) and secure remote copy (scp) were installed and tested to protect the cluster from unauthorized access and to encrypt all traffic between the cluster and client systems. The release of Kerberos ftp (kftp) at the time system implementation began was not large-file aware. It used a 32-bit address space for file size, limiting the size of any file that can be transferred to the cluster using the kftp utility to 2 GB in size. Solaris supports file sizes larger than 32-bits through the use of the large-file feature. Running the current release on the E10K systems in 64-bit mode will not solve the problem, as kftp is a 32-bit application, which does not use the source interfaces for a 64-bit file address space.

On a 32-bit system, a large file is a regular file whose size is greater than or equal to 2 GB (2**31 bytes). A small file is a regular file whose size is less than 2 GB. A utility is large file aware if it can process large files in the same manner as it does small files. A utility that is large file aware is able to handle large files as input and generate large files as output. Large-file aware programs can run in either 32-bit or 64-bit mode, and use the appropriate source interfaces for a 64-bit file address space. [5] Work performed both, locally and by NRL, resulted in a large-file aware kftp for the RMSS.

Direct user access to tapes is not required by the RMSS Cluster. SAM-QFS was the only application that requires tape device and tape access. SAM-QFS manages all user requests for data and provided the necessary interfaces and security.

2.3.7 Interactive Access

Support for required system services to support user requirements cannot be separated from the interactive login services. Each userid will have interactive access to the system. Solaris 7 does not support the restriction or limitation of userid interactive access.

2.3.8 STK HA-ACSLS Subsystem

The STK High Availability Automated Cartridge System Library Software (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

Clariion Application Transparent Failover (ATF) software. Ultra-10, LMU and Ultra-10 peripheral and network adapter failover is accomplished by Veritas Cluster Server (VCS).

2.3.9 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 then 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. One agent daemon runs for each configured resource type, and each type is unique to a particular hardware resource (persistent resources) or software resource. The core release package contains agents for predefined resource types and an environment for developing new agents for site-specific resource types. Custom-developed agent scripts were required to accomplish the failover of SAM-QFS. These scripts were developed during the configuration of VCS. Discussions held with the Aeronautical Systems Command (ASC) MSRC regarding these scripts have identified several important considerations with regard to SAM-QFS state at the time of failover.

The disk cache for all user data is deployed across the T3 FC-AL disk arrays. VCS requires that both E10K nodes in the RMSS Cluster have a path to the T3 disks. The Ancor FC switches provide these paths. The switch configuration is such that an alternate path to each T3 CU also exists. Should one E10K Host Bus Adapter (HBA) fail, the dynamic-multipathing (DMP) capability provided by Veritas Volume Manager is used to provide resilient access paths within the same cluster node. This configuration allows for the loss of one path to user disk cache or metadata without triggering a failover to the other node.

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

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, which includes the robotic arms inside LSMs where the tapes are housed. HA-ACSLS presents the control interface to clients to request tape mounts and dismounts of the robotics only. Data flow from the tapes is accomplished by a direct path (SCSI-3) between the SAM-QFS host, E10K, and the tape drives. 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. This software manages tape cartridge contents, generates requests for cartridges and transfers data to and from cartridges. The client software is not part of the Library Server product, but rather is incorporated in the SAM-QFS product.
The Library Server software resides on the server system and manages the storage and movement of tape cartridges and the use of library resources. It translates requests for tape cartridges, received from the client software, into cartridge movement requests for the LMU.

SAM-QFS communicates with HA-ACSLS for tape mounts only. It will request tape mounts for archive and stage operations, using its own daemon ssi_so. This daemon communicates with the HA-ACSLS Client System Interface (CSI) on the HA-ACSLS Ultra-10 system. The SAM-QFS ssi_so daemon is a shared object version of the SSI (Storage Server Interface) provided by STK. It is part of the base SAM-QFS software release. [1]

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 has special implications on the design of the VCS failover scripts, which must reestablish machine state.

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 the most significant of these performance tests.

3.1 Simple File Transfer Test

The simple file transfer test was an ftp transfer of compressible and uncompressable files from the original DMF File Server to the RMSS using the default network route for both systems. This route took files out of DMF via a Cray Gigaring FDDI adapter to a Cisco 1400 FDDI Hub. The files were then passed from the hub to a Cisco 7513 router where thry were switched onto an ATM OC-3 network. From the OC-3 network, the files were switched onto ATM OC-12c by a Cisco 8540 ATM Switch/Router to the RMSS ATM-622 network interface cards (NICs). Table 3-1, File Transfer Test Results, lists the results obtained by this test.

File Name	Transfer Time (Seconds)	Attribute	Bandwidth MB/sec	Size (Bytes)
1mb_ascii	0.22	compressable	4.55	1,048,576
1mb_binary	0.22	uncompressable	4.55	1,048,576
1gb_ascii	497	compressable	2.06	1,073,741,824
1gb_binary	261	uncompressable	3.92	1,073,741,824
5gb_binary	3901	uncompressable	1.31	5,368,709,120

Table 3-1

File Transfer Test Results

The 1 MB files transferred memory-to-memory, reporting completion before the files were flushed from memory buffers to disk. Therefore, these results measure the total network transfer time and latencies for a 1 MB data transfer. The 1 and 5 GB tests required extensive writes to disk before the transfer was reported as complete. This introduced disk latencies and I/O transfer bandwidths into the results.

3.2 File Compression Tests

The two 1 GB files used in the Data Transfer Test were compressed on one node of the RMSS Cluster. This was to measure the compressibility of the files using the Solaris implementation of the standard UNIX compress(1) command. The results are shown in Table 3-2, File Compression Test Results.

		Starting Size	**Ending Size**	**Comp r.**	**Time to Compress (Seconds)**		
File Name	**Attributes**	**(Bytes)**	**Bytes**	**Ratio**	**Real**	**User**	**System**
1gbascii	compressable	1,073,741,824	4,502,917	238:1	219.8	116.2	18.1
1gbbinary	uncompressable	1,073,741,824	1073741824	1:1	571.0	202.4	66.6

Table 3-2

File Compression Test Results

As expected, the binary file is shown to be uncompressible, an important attribute in later tests. The ASCII file is highly compressible, reducing from 1 GB to 4.3 MB, 238 times smaller than the original file. This, too, is an important attribute used in later tests.

3.3 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.

	File	**File Size**	**Time**			**Effective**
Operation	**Name**	**(Bytes)**	**Data Transfer**	**Tape Position**	**Total**	**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, consistient with the higher performance of 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. No other SAM-QFS actions involving the tape took place during this time, so that the tape was positioned at the write location for the "new" file when it was archived. The effective bandwidth (5 GB/Total Time) for both archive operations differed only by the positioning time in the first archive. 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. Compressibility increases the bandwidth dramatically. The effective bandwidth on recalls back to disk cache (stages) is far higher due to the higher read speeds of the tape drive. An effective bandwidth for recalls of 8.51 MB/sec was achieved. This bandwidth is consistient with average performance seen on the RMSS since transition and production usage began.

3.4 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 and Archive Server (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 number of files. In this case, the fixed upper file bound was 400 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. To increase fidelity to the production workload, binary, ASCII text, tar and compressed files were selected to recreate the nature of the data typical of the filesystems. The profile of the test data set created is listed in Table 3-4, Profiled Data Set File Distribution. All of these files together created a test set of nearly 6 GB in size (6,373,472,256 bytes).

File Size Range	/u/a	/u/b	/u/c	/u/d	Total
64KB	20	28	84	44	176
512 KB	16	16	24	28	84
1 MB	8	4	4	16	32
10 MB	4	24	36	8	72
50 MB	4	4	12		20
100 MB		4		4	8
500 MB		4			4
1 GB		4			4
Total Files	52	88	160	100	400

Table 3-4
Profiled Data Set File Distribution

The tests conducted with the profiled data set involved single and multiple stream transfers across a “direct” (via a Cisco 8540 ATM Switch/Router) ATM OC12c network. The source system was E10K with 64 CPUs and 64 MB memory. 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. In this test, four file streams, one for each planned production filesystem, were originated from the source system test. Two streams, one for filesystem /u/a and one for /u/c were transmitted to RMSS node 0. The two streams for filesystems /u/b and /u/d were transmitted to RMSS node 1. Data transfers across the ATM network were accompished using rcp. Elapsed time, from the first byte transferred to the last byte archived, was measured to determine an effective minimum bandwith of the cluster. At the time this test was conducted, the RMSS cluster was configured with two STK 9840 tape drives each. Archiving was supported for each filesystem by one tape drive.

Table 3-5, File Stream Network Transfer Timing Results, shows the results obtained for the network to disk cache data transfer portion of the test. Table 3-6, File Stream Archiving Timing Results, shows the results obtained for each stream as it was archived from disk cache to tape. Files that have been archived have transitioned through an important state change. Once a file has been archived, it is elligible to be released from disk cache and is considered a "permanent" member of the set of managed files. Thus, it is important to measure the elapsed time for files to transfer from a source system and become archived. The total time for each stream is listed in Table 3-7, End-To-End Stream Throughput. Table 3-8, Aggregate Stream Performance, restates this same timing information as a composite for all four streams combined.

Data Stream	Start Time	End Time	Elapsed Time (Seconds)	Bandwidth (MB/Sec)	Average Seconds/File	MB Transfered
/u/a	18:32:36	18:34:09	93	1.22	1.79	113.26
/u/b	18:32:36	18:39:23	407	11.95	4.63	4,862.19
/u/c	18:32:36	18:38:54	378	1.89	2.36	713.40
/u/d	18:32:36	18:37:27	291	1.34	2.50	389.29

Table 3-5
File Stream Network Transfer Timing Results

Achive Stream	Start Time	End Time	Elapsed Time (Seconds)	Bandwidth (MB/Sec)	Average Seconds/File	Files Archived
/u/a	18:35:55	18:44:25	510	0.22	9.8	52
/u/b	18:34:34	18:42:47	493	9.86	5.6	88
/u/c	18:34:52	18:45:32	640	1.11	4.0	160
/u/d	18:34:22	18:39:12	290	1.34	2.50	116

Table 3-6
File Stream Archiving Timing Results

End-to-End Streams	Start Time	End Time	Elapsed Time (Seconds)	Bandwidth (MB/Sec)	Average Seconds/File
/u/a	18:32:36	18:44:25	709	0.16	13.63
/u/b	18:32:36	18:42:47	611	7.96	6.94
/u/c	18:32:36	18:45:32	776	0.92	4.85
/u/d	18:32:36	18:39:12	396	0.98	3.41

Table 3-7

End-To-End Stream Throughput

Operation	Start Time	End Time	Elapsed Time (Seconds)	Bandwidth (MB/Sec)	Average Seconds/File
Transfer Data	18:32:36	18:39:23	407	14.93	0.98
Archive Data	18:34:22	18:45:32	670	9.07	1.61
End-to-End	18:32:36	18:45:32	776	7.83	1.86

Table 3-8

Aggregate Stream Performance

Judged by results, the RMSS can 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 tape drives for Node 0 and eight tape drives for Node 1, with the ability to increase these by a factor of three if required. Table 3-9, Projected System Performance, shows the test results as projected to a system with 15 total tape drives (7 drives on Node 0 and 8 drives on Node 1) and attained production ATM transfer rates. For the same amount of profiled data, using a 72% overlap between network to disk transfer time and disk to tape transfer time, such a system would accommodate the stream in 216 seconds, a bandwidth increase of 3.6 times.

Parameter	Test Scenario	Projected to Production System
Total Network Bandwidth	155 MB/sec	155 MB/sec
Network B/W Used	15 MB/sec (10%)	50 MB/sec (Observed)
Total Tape Bandwidth	40 MB/sec	150 MB/sec
Tape B/W Used	10 MB/sec (25%)	37.5 MB/sec (25%)
System Bandwidth	7.8 MB/sec	
Network Transfer Time	407 sec	122 sec
Archive Transfer Time	670 sec	178 sec
Total Stream Time	776 sec	(122+178)*.72 = 216 sec
Network/Archive Overlap %	776/(407+670)=72%	72%

Table 3-9

Projected System Performance

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 per 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.

3.5 Two Million Files Tests

These tests were conducted to measure critical performance metrics for large filesystems. The production RMSS was designed to handle very large file counts within a single filesystem, making the time to delete and scan directory entries important. The type of scanning measured is the time required to scan all directory entries in the inode metadata file when SAM-QFS is first started. During this interval, files are available to the users, but no new files can be archived. This interval can prove critical if large numbers and sizes of files were created just prior to the last time SAM-QFS was shutdown or during the time that the scan is taking place. The goal is to control new file generation and file modification to prevent overflow of disk cache until the files can be archived and their space in disk cache released for new files. A filesystem of 2 million files, 1000 directories, and 107 other files was created. Each terminal node directory entry contained 2000 files. The results are presented in Table 3-10, Two-Million File Test Results.

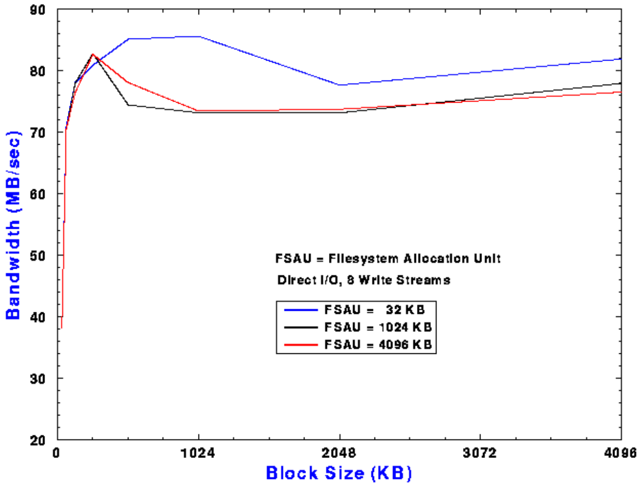
Operation	Elapsed Seconds	Files/Second
Delete	4,751	421
Scan	410	4,881

Table 3-10
Two-Million File Test Results

3.6 Disk Performance Tests on Nobles

Extensive testing of early T3 Control Units was performed and repeatedly achieved peak write-performance of approximately 86 MB/sec using direct I/O on transfers of 2 MB blocksize with four read-write streams and disk allocation units in excess of 1 MB under SAM-QFS/Solaris 2.6 to a RAID-5 configuration. Using the same configuration and transfers under a Unix File System (UFS), peak write performance of 56 MB/sec for 2 MB blocksizes and 60 MB/sec on 8 MB/sec was consistently attained. The performance rolloff was minimal for 1 MB+ disk allocation units. Performance ramp-up for blocksizes under 1 MB was steep. Figure 3-1, MaxStrat Noble (SUN T3) Write Performance, shows the performance measurements (bandwidth) taken for filesystem allocation blocksizes (FSAU) of 32 KB, 1024 KB and 4096 KB for I/O blocksizes ranging from 32 KB to 4096 KB.

Figure 3-1
SUN T3 Write Performance



4 Transition to Production and Future Work

Moving over 200 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. Three filesystems on MSAS1 must be transferred into four filesystems on the RMSS. This process is anticipated to require approximately one year to complete. 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. It is fairly easy to replicate the existingdirectories and files on the new host server while retaining some level of hierarchical storage content, but the cost of providing the support for "DMF translation", tape I/O management, and other tasks are high risk elements.

Various other alternatives for data archive migration were considered. These included Recalling and converting over 200 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 project, by account number, etc. All were determined to be technically impractical and administratively complex on a scale as large as 200 TB. After considering all of these alternative, 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 chage the reading (staging) a file from a CDROM, to reading a file using a simple "rsh" across a network.

The process used for data transition is illustrated in Figure 4-1, Transition Data Movement. 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 communcating 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.

Once the file had been rearchived, it is 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.

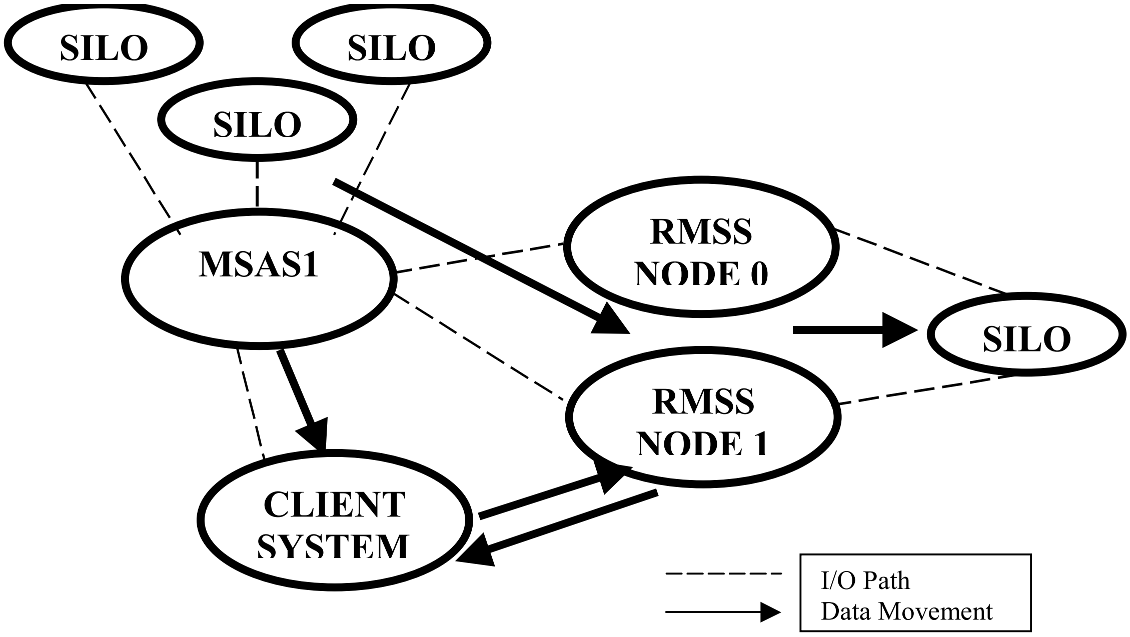


Figure 4-1
Transition Data Movement

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 (or could be new) and is no longer needed on the old archive. This greatly simplfies 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. As of mid-February, 2001, 33% of the data in this filesystem (total of 60 TB) had been moved to the RMSS system. 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 (135 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.

4.2 Future Work

The transition of the RMSS to full production must be completed before new technology can be incorporated into the server. 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. Several avenues of study are progressing to stay at the forefront of mass storage technology. These are:

- Fibre Channel Full Fabric Support (SAN)

  The purpose is to determine the best means to deploy SAN technology to support the MSRC and its mass storage requirements.

- Disk Technology

  Invesitgations into future disk technology is being conducted. The disk cache is a critical component in the performance of the HSM server and faster disks with greater capacity are an important continuing study.

- Software Technology

  A re-engineered version of the SAM-QFS softwear is now available. This version represents an important step forward in the HSM software technology at the heart of the RMSS and provides important advances in robustness. The need to maintain a stable environment for the RMSS during transition presently precludes installation of this software.

- Tape Technology

  Fibre Channel tape technology brings the ability to fail-over tape drives. This maintains the I/O bandwidth critical to continued operations of the RMSS. In addition, capacity tape-drives, such as the STK 9940 are being investigated for potential inclusion into the RMSS resource complement at a future date.

5 Conclusion

The MSRC RMSS is proving itself capable of handling the workloads it was designed to handle and is flexible enough to scale to handle its projected requirements. The success of the RMSS is also measured in the fact that the other two MSRCs are also 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 exented 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