HPSS at Los Alamos: Experiences and Analysis

Per Lysne
Gary Lee
Lynn Jones
Mark Roschke
Los Alamos National Laboratory, Los Alamos, NM

Abstract

The High Performance Storage System (HPSS) is currently deployed on the open and secure networks at Los Alamos National Laboratory (LANL). Users of the Accelerated Strategic Computing Initiative (ASCI) system with 6,144 processors and our similar Advanced Computing Laboratory (ACL) system, both from SGI/Cray, access HPSS for their data storage. We discuss our current HPSS configurations and how our users access HPSS. We analyze the performance between HPSS and these systems. We also discuss our projected storage and storage performance requirements for the next several years and what we are planning to meet those needs.

Introduction

HPSS is currently deployed by the Computing Group at LANL as the primary archival storage facility for users of the ASCI computers on both the open and secure networks. We begin with a brief description of the computing and storage environments at LANL followed by an overview of HPSS itself. This is followed by an analysis of the future ASCI storage needs at LANL. Our current equipment and configuration are then described, and our plans to meet our future requirements are outlined. At this point we describe the HPSS access methods provided to users at LANL and then move on to discuss in detail the usage patterns and performance seen on our current HPSS systems. Finally, we conclude with some experience gained setting up and operating our production HPSS systems.

LANL Computing and Storage Environment

The computing environment at LANL is partitioned into separate networks for unclassified (open) and classified (secure) work. The primary worker machine on each network is an SGI/Cray Origin 2000. As shown in Table 1, the open Advanced Computing Laboratory system is configured into nodes of n x 32 MIPS R10K processors, where n=1-4, for a total of 768 processors and 192 GB of memory. As shown in Table 2, the secure ASCI Blue Mountain system is configured into 48 nodes of 128 MIPS R10K processors each for a total of 6,144 processors and 1.5 TB of memory. These configurations change periodically when new equipment is added or when nodes are split or merged.

Both open and secure networks utilize a HIPPI infrastructure for data transfer. As shown in Figure 1, each node in the open system has between one and four HIPPI connections to an internal switch fabric and one connection for external network services, such as HPSS data transfer. Each node in the secure system has twelve HIPPI connections to an internal switch and one connection for external network services. External FDDI connections are provided on both networks for additional network services such as HPSS control information.

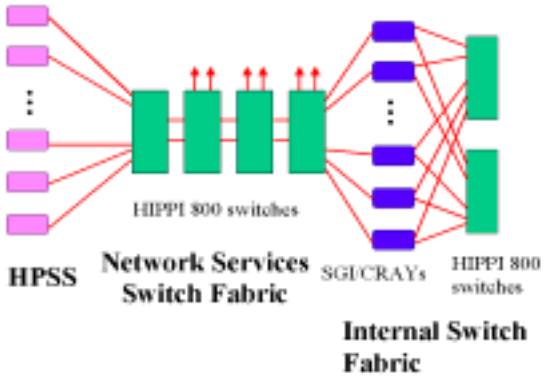

Figure 1. LANL HIPPI 800 Network

Most worker machines on the LANL networks are configured with a substantial amount of direct-attached disk storage. These systems are used for large scale physics, climate, and other types of modeling, and these codes either run completely in system memory or use direct-attached disk for run time storage.

Each LANL network contains an HPSS system as well. The HPSS systems are intended mainly for archival storage of large files, such as the output from modeling codes. A general configuration diagram which applies to both LANL HPSS systems is shown in Figure 2. The current equipment deployed in each system is given in Tables 6 and 7. Capacity and usage information is shown in Table 3.

Other storage systems and interfaces, such as IBM's ADSM, TransArc's DSF, and NFS are also provided on the LANL networks for such tasks as workstation backups and storage of small files.

Table 1. Open SGI/CRAY Computing System

Model	Quan	# CPUs	CPU Speed	Memory
Origin 200*	3	2	180 MHz	256 MB
Origin 2000	2	32	195 MHz	8 GB
Origin 2000	1	32	250 MHz	16 GB
Origin 2000	1	32	250 MHz	32 GB
Origin 2000	1	64	195 MHz	16 GB
Origin 2000	5	128	195 MHz	32 GB
* Front-end system				

Table 2. Secure SGI/CRAY Computing System

Model	Quan	# CPUs	CPU Speed	Memory
Origin 200*	2	2	180 MHz	256 MB
Origin 2000	48	128	250 MHz	32 GB
Onyx 2**	1	8	195 MHz	4 GB
* Front-end system		** Visualization server		

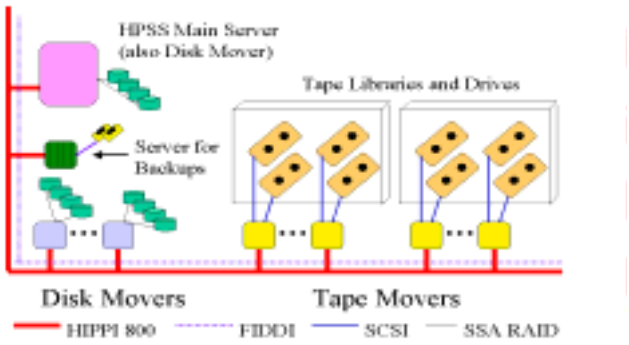

Figure 2. Los Alamos HPSS Configuration

Overview of HPSS

HPSS is a highly-scalable, parallel, high-performance hierarchical storage management software system. It is being developed by a collaboration involving IBM and four US Department of Energy (DOE) laboratories (Los Alamos, Lawrence Livermore, Oak Ridge, and Sandia National Laboratories). The ASCI program funds a large part of HPSS development, but HPSS has been available as a product from IBM Global Government Industries since 1996. In 1997 HPSS received an R&D 100 Award from R&D Magazine for its scalable architecture, network centric design which supports direct network attached devices, and parallel I/O capabilities.

HPSS has been developed to meet the need for higher performance and larger capacity data storage systems to be used in high-performance computing environments. HPSS is designed to store millions of files, petabytes (10^{15}) of data, and to transfer gigabytes (10^{9}) of data per second using parallelism and network connected storage devices.

As part of its network-centered design, the functionality of the HPSS system is partitioned among many independent servers. The HPSS servers and data movers can be distributed between different machines on a high performance network to provide scalability and parallelism. Actual data transfers occur directly between the client and the device controlling the storage. This may be done using third-party protocols, such as IPI-3, or with TCP/IP. The controller may be intelligent (e.g. Maximum Strategy Disk Array), or may be a low-cost Unix processor, or Protocol Engine, executing HPSS Mover code. Multiple, parallel movers can be used in a single transfer operation to increase throughput. For flexibility and performance, HPSS also allows the use of separate networks for control information and data transfer.

HPSS has been developed for Unix systems and requires no kernel modifications. It was originally developed for AIX, but ports of the HPSS Mover and Client API are in progress for Ultrix and Irix as well as a full port to Solaris. The OSF Distributed Computing Environment (DCE) and TransArc's Encina are the basis of HPSS's distributed, transaction-based architecture. HPSS also uses the DCE Remote Procedure Call (RPC) mechanism directly for non-transactional RPCs, and the DCE Security and Cell Directory Services. Along with Encina for transaction management, HPSS depends on the Encina Structured File Server (SFS) as its metadata manager.

HPSS supports storage hierarchies. A storage hierarchy consists of multiple levels of different types of storage. An HPSS system commonly supports several hierarchies with different service characteristics, such as access time, maximum file size, number of data copies, and data transfer rate. At the user interface, these hierarchies are called classes of service. When a user stores a file to HPSS, a default hierarchy is selected according to the file size and other information available to HPSS depending on the interface being used. Depending on the interface they are using, the user may also select a particular class of service based on their own requirements.

Files move up or down the hierarchies via migrate and stage operations which are controlled by site configured policies on the basis of file usage and storage availability. Hierarchies are commonly configured with faster, more expensive types of storage at the top, and slower, less expensive storage at the bottom in order to achieve a caching benefit. For instance, a storage hierarchy may consist of disk storage followed by one or more levels of tape storage. Large files that are being archived may be written directly to a tape-only hierarchy for reduced cost and better performance. Multiple data copies are currently supported in HPSS by configuring a storage hierarchy with disk at the top followed by two levels of

tape below. Any file written to this hierarchy will then be copied onto both levels of tape when migration occurs.

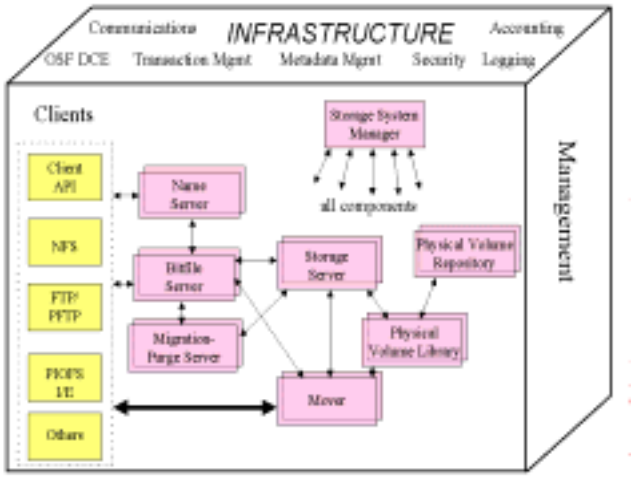

Figure 3. HPSS Infrastructure

The HPSS servers currently include the Name, Bitfile, Migration/Purge, and Storage Servers, as well as the Physical Volume Library and Repository, Mover, and Storage System Manager. In the next release (4v1), HPSS will also utilize a Location Server which allows the Name Servers of multiple geographically distributed HPSS systems to present a common federated name space to the systems' users. Many of the servers may be replicated for improved performance. Details on each of the HPSS servers may be found in the references [1,2].

Figure 3 shows the HPSS infrastructure components along the top. The HPSS user interfaces, or clients, are shown along the left side of the same figure.

LANL HPSS Requirements

The requirements placed on the open LANL HPSS system are based on an extrapolation of current system usage. The current ACL system is capable of around 300 GOps. A 1.0 TOp machine is scheduled to be installed in 1999. This approximately represents a factor of three increase in computing power. Table 3 gives the open HPSS usage at the beginning and the end of 1998. The projected usage at the end of 1999 is found by multiplying growth that occurred during 1998 by the three increase in machine size.

The requirements on the secure LANL HPSS system are currently driven by the ASCI project. The ASCI machines at LANL are projected to reach 3-TOps in 1999, 10-TOps in 2000, and 30-TOps in 2001 or 2002. Since these machines are being used to run modeling codes with large output files, the requirements placed on HPSS are focused on these large files. In particular, ASCI has set out requirements for the aggregate HPSS system throughput, and also for the total HPSS system capacity. As of now there is no requirement for HPSS to provide multiple copies of stored files or any type of vaulting. There is an informal expectation that the LANL HPSS systems will support smaller files at some reasonable level. Table 3 gives the secure HPSS usage at the beginning and end of 1998.

Table 3. HPSS Usage and Capacity

Open HPSS			
	Jan '98	**Jan '99**	**Jan '00**
Usage	14.3 TB	30.5 TB	79.1 TB
Total Capacity	43 TB	43 TB	243 TB
Secure HPSS			
	Jan '98	**Jan '99**	**Jan '00**
Usage	2.3 TB	18.5 TB	2 PB
Total Capacity	31 TB	31 TB	2.31 PB
Jan '00 figures are projections			

Two models are being used to project the future secure HPSS system requirements at LANL. The first model, the Data and Visualization Corridor (DVC) model is the product of a series of workshops jointly sponsored by the DOE and the National Science Foundation (NSF) in the spring of 1998 [2]. The second model, the ASCI Data Storage Curve, is a series of projections generated by program managers and others from the national laboratories and DOE.

The DVC model is based on an extrapolation of current trends in high performance computing. The plans and requirements of ASCI are included in this model. As shown in Table 4, the DVC model estimates that every TOp of machine performance is capable of generating 1-2 GB/s of I/O to direct attached storage. Based on past system usage, the DVC model states that the system area network (SAN) need only support about one-tenth of the direct-attached I/O rate. Past experience indicates that the SAN is actually capable of about one-third of it's rated performance, with the other two-thirds being lost in the operating systems and network protocols involved. For this reason, the DVC model calls for the SAN to be scaled at three times the theoretically required capacity, or approximately one-third of the direct-attached machine I/O rate. The SAN capacity gives the sustainable maximum throughput rate which is being used to size the LANL HPSS systems. Table 4 also shows how the model anticipates this throughput requirement to be met by a SAN composed of sixteen parallel 100 MB/s HIPPI-800 links in 1999, and with the same number of 800 MB/s HIPPI-6400 links in 2001.

The ASCI Data Storage Curve gives projections for the required total storage system capacity. Table 5 gives the total required capacities for future HPSS systems at LANL.

Thus far the actual HPSS usage has been substantially below what has been estimated. We believe this is

due to the lag between the time when each successive ASCI machine is available and the time when the modeling codes are ready for production runs on those machines. Realistically, this means that the current LANL storage requirements may be adjusted downward sometime in the future.

Table 4. DVC Model/Throughput

	1999	2000	2001
Machine Size	3 TOp	10 TOp	30 TOp
Direct Attached I/O	3-6 GB/s	10-20 GB/s	30-60 GB/s
SAN I/O	1.6 GB/s (16*HIP PI-800)	7.2 GB/s (16*HIP PI-6400)	12.8 GB/s (?)

Table 5. ASCI Storage Curve/Total Capacity

	1999	2000	2001
Total System Capacity	2 PB	3.5 PB	5 PB

LANL HPSS Hardware and Configuration

The HPSS software in both the open and secure systems is configured in much the same way. Both systems support several storage hierarchies with disk at the top and two shared levels of tape below. These disk-dual tape hierarchies are configured to accommodate relatively small files of different sizes. The only difference between these hierarchies is the size of the space allocation unit on disk. Hierarchies intended for smaller files have a smaller allocation unit, and hierarchies for larger files have a larger allocation unit. This type of configuration reduces wastage on disk. All of these small-file hierarchies share a common pool of tape below on which multiple copies are maintained via migration. Although there is no formal requirement for multiple copies, our policy to date has been to make such copies whenever the cost is low. Both systems also support direct tape storage, which is selected automatically for files above a given size, but which can also be selected by the user for files of any size. Both systems support a 1-way stripe tape hierarchy, and the secure system also supports a 2-way tape stripe. The current release of HPSS (3v2) does not support multiple copies for direct tape storage.

The current hardware included in the open LANL HPSS system is given in Table 6. The current secure equipment is given in Table 7.

The open hardware is configured such that the SSA RAID disk is distributed equally among the three disk mover machines. This disk is shared by four storage hierarchies which utilize disk as the top level. The tapes are shared among all five hierarchies which use tape, with two of the 3590 drives being connected to each of the six tape mover machines. Each 3494 library contains six 3590 tape drives.

Table 6. Current Open HPSS Equipment

Function	Quan	Description
HPSS & SFS Server	1	IBM RS6000-R24
Disk Movers	3	IBM RS6000-43P-240
Disk Drives	127GB	IBM SSA RAID
Tape Movers	6	IBM RS6000-43P-140
Tape Libraries	2	IBM 3494
Tape Drives	12	IBM 3590

Table 7. Current Secure HPSS Equipment

Function	Quan	Description
HPSS & SFS Server	1	IBM RS6000-H50
Disk Movers	1	IBM RS6000-R24
Disk Drives	127GB	IBM SSA RAID
Tape Movers	6 1	IBM RS6000-43P-140 IBM RS6000-R24
Tape Libraries	2 1	IBM 3494 STK Powderhorn
Tape Drives	13 5	IBM 3590 STK Timberline

The secure disk is all directly attached to the single disk mover and is shared between four hierarchies. Two 3590 drives are connected to each of the six 43P tape movers, except for one mover which supports three drives. All of the Timberline tapes are connected to the R24 tape mover. In the secure system, the Timberline tapes are used to support the disk hierarchies because of these tape's limited 800 MB capacity. The 3590 tapes are used in support of the 1-way and 2-way stripe direct tape hierarchies.

For 1999, equipment purchases are planned which will bring the open and secure systems to the levels given in Tables 8 and 9. Some equipment will also be retired during this process.

Table 8. 1999 Open HPSS Equipment

Function	Quan	Description
HPSS & SFS Server	1	IBM RS6000-H50, 4CPU
Disk Movers	4	IBM RS6000-43P-240
Disk Drives	1 TB	Fibre Channel RAID
Tape Movers	16	IBM RS6000-43P-140
Tape Libraries	2 1	IBM 3494 STK Powderhorn
Tape Drives	12 20	IBM 3590 STK Eagle

The open equipment purchases are driven roughly by multiplying the current equipment by the factor of three increase in capability of the open ACL machine. The secure equipment is driven by the 1.6 GB/s through-

put requirement and by the required 2 PB of total capacity.

The 1.6 GB/s aggregate throughput will be achieved by way of 32 four-way tape stripes, which we anticipate will each achieve 50 MB/s. Each Eagle tape cartridge has a raw capacity of 20 GB. We have observed a compression ratio of about 1.6, so after compression each cartridge holds approximately 32 GB. At this ratio, 62,500 cartridges are required to hold 2 PB, and each STK silo holds approximately 6000 cartridges. Hence, the ten silos will nearly meet our 2 PB requirement.

Table 9. 1999 Secure HPSS Equipment

Function	Quan	Description
HPSS & SFS Server	1	IBM RS6000-H50, 4CPU
Disk Movers	1 4	IBM RS6000-43P-240 IBM RS6000-43P-140
Disk Drives	1.5 TB	Fibre Channel RAID
Tape Movers	70	IBM RS6000-43P-140
Tape Libraries	2 10	IBM 3494 STK Powderhorn
Tape Drives	13 128	IBM 3590 STK Eagle

Equipment for ensuing years will be sized in the same way. Based on the trend towards large numbers of tape drives, libraries, and movers demonstrated by the secure 1999 system, increased capacity equipment is clearly desirable. Without larger capacity, higher-bandwidth tape equipment in the future, the amount of equipment needed to support the future system requirements will quickly become unmanageable. Also, the maximum bandwidth available for a single file transfer from tape is determined by the width of the tape stripe. HPSS does not support any type of redundancy on tape stripes, so we are currently unwilling to go beyond a stripe width of four. In the future we hope to be able to purchase Redundant Arrays of Inexpensive Tape (RAIT) equipment to solve this problem and enable us to use wider stripes.

Access Methods

HPSS supports several user clients or interfaces. Some are industry standards, such as FTP and NFS version 2. HPSS includes a client API, which is a programming interface that allows direct access to HPSS. HPSS can also act as an external file system for the IBM SP Parallel I/O File System. Additional interfaces available in the next release (4v1, 12/98) are for the Distributed File System (DFS) and MPI/IO. The most important interfaces used at LANL, however, are Parallel FTP (PFTP) and the Parallel Storage Interface (PSI).

PFTP is a modification of standard FTP which has been developed to work with HPSS. It supports the standard FTP commands as well extensions to increase data transfer performance by allowing data to be transferred in multiple parallel streams. Parameters such as stripe width and block size are set automatically, but can also be controlled by the user. For each file stored, PFTP allows the user to provide the file size in advance to the HPSS system. HPSS then uses this hint to select the class of service which best meets the user's needs.

The PSI interface has been developed locally at LANL. This interface runs directly on top of PFTP and provides a number of useful features. PSI has a Unix-like command set which provides the user with a familiar set of tools to manipulate HPSS files. PSI provides an automatic retry mechanism for file transfers, and a trash-can mechanism which reduces the likelihood of inadvertent data loss. A test facility allows a user to see the actions that would be taken on complex commands, including subtree copies or modifications, before committing to such commands. User commands and the performance statistics for these commands are logged to aid in debugging and tuning the HPSS system. Additional performance increases through PSI are being pursued by way of new features such as simultaneous file transfers, optimizing the sequence of file transfers, and with conditional file transfers.

HPSS Usage Statistics And Performance

The LANL HPSS systems came on line near the end of 1996. The systems were operated in friendly-user mode for a year, and then moved into production at the end of 1997. Prior to entering production, approximately 8 TB of data from another LANL mass storage system, HPDS, were migrated to the open HPSS. At present, the open system contains 721,976 files for a total of 20.7 TB of user data. The secure system contains 510,001 files for a total of 15.7 TB. Both systems make multiple copies of files stored in disk-tape hierarchies. Taking these copies into account, the open system contains a total of 30.5 TB and the secure 18.5 TB. Current growth in both the open and secure systems, counting copies, is about 2 TB per month.

The distribution of file sizes on both systems is weighted heavily towards small files. File size data is collected by histogramming all of the file sizes on a 10KB basis. Two of the curves in Figure 3 show the cumulation of files by size (i.e. the percentage of files below a certain size). This figure shows that nearly 90 percent of the files on both systems are below 50 MB. Figure 4 also shows the cumulation of data (i.e. the percentage of data contained in files below a certain size). These curves show that 90 percent of the data in the open system is contained in files less than 15 MB, but the 90 percent mark in the secure system is not reached until nearly 43 MB. This demonstrates a trend towards larger files on the secure

system, which may be attributed to an ongoing effort to educate the secure users about the performance benefits of large files.

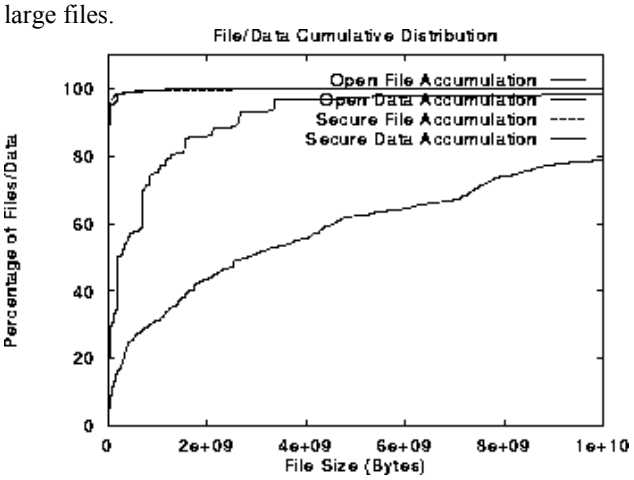

Figure 4. File/Data Cumulation

Table 10. Distribution of Files and Data

Open HPSS System		
	Files	**Data (GB)**
Total:	721,976 (100%)	20,665 (100%)
1 Day:	25,275 (3.5%)	92 (0.4%)
1 Week:	34,137 (4.7%)	702 (3.4%)
1 Month:	54,068 (7.5%)	1,310 (6.3%)
1 Year:	471,365(65.3%)	9,992 (48.4%)
Longer:	137,131(18.9%)	8,569 (41.5%)
Secure HPSS System		
	Files	**Data (GB)**
Total:	510,001 (100%)	15,693 (100%)
1 Day:	2,193 (0.4%)	391 (2.5%)
1 Week:	28,763 (5.6%)	620 (4.0%)
1 Month:	115,859(22.7%)	4,377 (27.9%)
1 Year:	360,644(70.7%)	9,609 (61.2%)
Longer:	2,542 (0.5%)	696 (4.4%)

Table 10 shows the distribution of files and data on the LANL HPSS systems by the time of last access. This figure shows that 19 percent of the open files have been inactive for over one year. These files constitute 42 percent of the system data however (not counting multiple copies). Most of these files are probably the output of large models which were migrated from HPDS prior to the open HPSS going into production. Table 10 shows that less than one percent of the files on the secure system have been inactive for more than a year, and only total 4.4 percent of the data. If files which have been accessed within the last month are considered active, then 15.7 percent of the open files, or 10.1 percent of the data is active. By the same measure, 28.7 percent of the secure files and 34.4 percent of the secure data are active. Over the long term we expect most of the data on our HPSS systems to be inactive, but the systems have not been in production long enough to determine this.

On the open system 5.1 percent of the files are stored to direct-tape hierarchies. Not counting multiple copies, this represents 52.3 percent of the open data. In the secure, only 6.9 percent of the files are on direct tape, but this accounts for 82.1 percent of the data. Again, this may be explained by our efforts to encourage the secure users to store large files, which by default go directly to tape. Also, the secure users have been encouraged to store small files to tape when rapid access to these files is not necessary.

Our performance testing methodology has been fourfold. First, we determined the best possible performance between the client machines and HPSS using ttcp tests. This gives us a measure of our HIPPI network's throughput as well as the I/O capabilities of the client machines involved. The results are shown in Table 11.

Table 11. TTCP (Memory to Memory) Tests

Description of Test	**Result (Single Stream)**	**Result (Dual Stream)**
Between SGI/Cray Nodes	71 MB/s	NA
Write to 43P Mover	20 MB/s	24 MB/s
Read from 43P Mover	31 MB/s	27 MB/s
Write to 43P Mover (2 CPU)	NA	32 MB/s
Read from 43P Mover (2CPU)	NA	32 MB/s
Note: Other system was SGI/Cray or Cray M98		

Due to anomalies with protocol stack processing in AIX and Irix, we expected results to differ widely depending on the buffer and window sizes. As an example, when running the ttcp utility between an SGI and an AIX machine, the transmission rate dropped from 30 MB/s to 200 KB/s just by changing the socket size on the transmitting socket from 60KB to 64KB. If the TCP_NODELAY option is added, the transfer rate drops instead from 30 MB/s to 10 MB/s. While we found many combinations that yielded poor performance, we also found a number of combinations that yield the performance given in Table 11. The best single stream result, 31 MB/s with the SGI/Cray as the sink and the AIX mover as the source, was 20% higher than any of the other single stream results.

In the past, several of our machines have used IBM HIPPI Microchannel network adapter devices (MCA). These adapters have their own peculiar configuration requirements, which differ depending on how the network is being used. For our purposes, where control is over FDDI and data over HIPPI, the FDDI network handles the small packets so the HIPPI network need only be optimized for large packets. In addition to the general network parameters and considerations, the MCAs have

buffer configuration issues to be explored. Although we delved into this in detail in order to avoid errors on the HIPPI network during congested periods, we have recently reconfigured our systems so that the bulk of data from our mover machines flows over Essential PCI HIPPI Adapters. These have a much smaller set of configuration options, and hence less room for error. The Essential adapters currently handle sustained writes of 20 MB/s and reads of 31 MB/s. The results in Table 11 also indicate that using a dual-CPU mover machine does not result in a 2x increase in throughput because the system appears to be constrained by the system bus at 32 MB/s.

Our second form of testing has been done to determine the raw transfer rates of the disk and tape devices used in our HPSS systems. These tests are designed to isolate the storage devices themselves to eliminate any effects the network or HPSS software may have on performance. At LANL we are using IBM SSA RAID disks for our disk storage. This has required careful testing and evaluation of parameter settings effecting the SSA RAID adapter. The adapter technology has been steadily improving over the time we have worked with it. Better adapters, microcode, and drivers have given improved RAID disk performance over the past two years. Initially, we were able to transfer data at approximately 5 MB/s to our RAID disk systems, but that has now improved to 20 MB/s for writes and 34 MB/s for reads. We performed similar tests on our IBM 3590 tape drives, and the results are given in Figure 12.

The device tests in Table 12 provide us with theoretically the best performance we should expect. Looking at the results from Tables 11 and 12, we expect that a single CPU 43P mover with an Essential PCI HIPPI adapter will be able to support two IBM 3590 tape drives running at their maximum rates for reads but will become a bottleneck for writes. The same mover can support a single SSA RAID subsystem for writes, but will become a bottleneck on reads.

Table 12. Device Tests

Description	Result
Writing to SSA RAID*	20 MB/s
Reading from SSA RAID*	34 MB/s
Writing to 3590 Tape	13 MB/s
Reading from 3590 Tape	11 MB/s
* Logical volumes are striped 4-way with one parity disk	

Our third test was performed on the HPSS software itself to determine the overhead associated with the transaction and metadata management versus the actual data transfer component. The 0-Byte create test simply creates files in the HPSS system of zero length. This tests the namespace operations necessary to create a file, but none of the storage allocation functions. The 1-Byte create test does the same thing, but also includes the storage allocation. The delete test simply determines the rate at which files can be deleted. These tests were run on our open HPSS system using our original IBM 580 server and also using our current IBM R24 server. As expected, the results in Table 13 show that improvements in the HPSS server performance result in better transaction performance and increases the rate at which files can be created and deleted. This table also shows the benefit of removing the disk mover function from the main server.

Table 13. HPSS Infrastructure Tests

Description	Create 0-Byte File	Create 1-Byte File	Delete File
IBM 580 Server (Disk Mover on Server)	2.5-3.5 Files/s	.75-.90 Files/s	2.3-3.4 Files/s
IBM R24 Server (Disk Mover on Server)	4.6-5.8 Files/s	1.6-1.9 Files/s	5.2-5.8 Files/s
IBM R24 Server (No Disk Mover)	5.0-5.8 Files/s	2.0-2.4 Files/s	6.0-6.6 Files/s

Finally, tests were performed between the SGI/Cray client and HPSS using PFTP. The results are shown in Table 14. These results show that HPSS disk performance is better with large files than small files. With small files, the speed of the metadata functions becomes the dominant factor. HPSS performance using tape is quite good for large files. This indicates that large archive files should be written directly to tape, rather than to disk and then migrated later to tape. HPSS write performance to disk is about 25% of the theoretical maximum. These tests were run using an SSA RAID adapter for the PCI bus.

Table 14. Client to HPSS Tests

Op	Source	Sink	File size	Result
Write	Client Mem	HPSS Tape	1 GB	11.40 MB/s
Read	HPSS Tape	Client Mem	1 GB	11.59 MB/s
Write	Client Disk	HPSS Tape	1 GB	11.00 MB/s
Read	HPSS Tape	Client Disk	1 GB	11.65 MB/s
Write	Client Disk	HPSS Disk	500 MB	6.60 MB/s
Read	HPSS Disk	Client Disk	500 MB	16.37 MB/s
Write	Client Disk	HPSS Disk	50 MB	5.60 MB/s
Read	HPSS Disk	Client Disk	50 MB	15.60 MB/s
Write	Client Disk	HPSS Disk	121 500B files	750 B/s (1.5 files/s)
Read	HPSS Disk	Client Disk	121 500B files	1250 B/s (2.5 files/s)

A different kind of issue that effects HPSS performance is how the various Classes of Service are configured. Users store a file to one of several classes of services,

each class of service having different characteristics such as file size, storage medium, and access time. A class of service must be set up correctly so that files are stored efficiently. Although the HPSS Administrative Manual addresses this, administrators have been known to set up a class of service poorly, and get surprisingly poor transmission performance. As an example, if the maximum allocation unit is set to 1 MB, and the maximum file size is set to 500 MB, there could be up to 500 units allocated for the file. This leads to excessive metadata overhead, which is all incurred during data transfer. The configuration of Classes of Service must be clearly understood in order to get the best performance out of an HPSS system.

Operational Experiences

Since going online in late 1996, the open and secure HPSS systems have been available over 95% of the time. During the past year there were several months without any downtime. These numbers are measured for the period between 8:00 AM and 8:00 PM seven days a week when the HPSS systems are required to be available. Scheduled downtime between these times is also not counted. During this time users of the secure system have been somewhat dissatisfied with the levels of performance delivered by HPSS. This has mainly been due to lack of hardware funding to support a sufficiently large system.

To date HPSS has been able to meet its requirements at LANL and is regarded as a success. We have encountered problems along the way which include stability of the HIPPI network and performance of the SSA RAID disks. The SFS database has had performance problems and has also proven particularly difficult to administer. DCE has suffered a limited number of problems, but Encina has performed quite well. The remainder of our hardware, and the RS6000s in particular, has proven quite satisfactory.

The HPSS software itself has had a few problems, but nothing insurmountable. The main complaint with HPSS itself has been the large operations and administration staff required to keep it in production, and also the non- intuitive nature of it's interface. The most successful HPSS deployments have been at HPSS development sites. Several sites without local development experience have found HPSS deployment a frustrating experience. A major focus of the next HPSS release (4v2, 2000) is to improve the manageability of the system.

Summary

HPSS has been successfully deployed in production status at LANL for over a year. At present, we are considering options for migrating data from an older data storage system, the Common File System (CFS), to HPSS. Meeting the considerable future requirements of the ASCI program continues to be a challenge. It will be necessary to continuously improve the HPSS software and to tune our HPSS hardware to achieve the maximum possible performance. High performance storage at LANL continues to be dependent on the development of new storage technology, such as higher capacity tape cartridges and drives, RAIT, network attached storage, and third party transfer equipment. Finally, the future success of HPSS at LANL will depend on the ASCI users structuring their codes and other procedures in such a way to take advantage of HPSS's capabilities.

References

[1] **HPSS System Administration Guide**

[2] R. W. Watson and Robert A. Coyne. **The Parallel I/O Architecture of the High-Performance Storage System (HPSS).** Proc. 14th IEEE Symp Mass Storage Systems. April 1995.

[3] Paul H. Smith and John van Rosendale. **Data and Visualization Corridors: Report on the 1998 DVC Workshop Series.** Caltech Technical Report CACR-164. September 1998.