Towards Petabytes of High Performance Storage at Los Alamos

Gary Lee (rgl@lanl.gov), Gary Grider (ggrider@lanl.gov),
Mark Roschke (mar@lanl.gov), and Lynn Jones (lkj@lanl.gov)
Data Storage Systems Group
Los Alamos National Laboratory
Los Alamos, New Mexico 87545 USA
Phone +1-505-667-2828, FAX +1-505-667-0168

ABSTRACT: The High Performance Storage System (HPSS) is currently deployed in the open and secure networks at Los Alamos National Laboratory. Users of the Accelerated Strategic Computing Initiative (ASCI) Blue Mountain MPP system and the Advanced Computing Laboratory's MPP system, both from SGI/Cray, access HPSS for their storage. We discuss our current HPSS configurations, how our MPP users access HPSS, and the performance between HPSS and the MPP systems. We will also discuss our projected storage and storage performance requirements for the next five to seven years.

Introduction

The High Performance Storage System is currently deployed by the Data Storage Systems Group at Los Alamos National Laboratory as the primary archival storage facility for users of the Accelerated Strategic Computing Initiative computers. Following a brief discussion of HPSS, we describe the configuration and status of HPSS and the SGI/Crays in the Laboratory's High Performance Computing Environment. Some performance measurements for HPSS are described in detail, along with some ideas for improving performance characteristics. We conclude with our projected ASCI storage and performance requirements.

Overview of HPSS

HPSS is a highly-scaleable parallel high-performance software system for hierarchical storage management ([1], [2]). It is being developed by a collaboration involving IBM and four US Department of Energy (DOE) laboratories (Los Alamos National Laboratory, Lawrence Livermore National Laboratory, Oak Ridge National Laboratory, and Sandia National Laboratory Albuquerque). HPSS has been available as a service offering from IBM Global Government Industries for almost two years. ASCI funds a large part of HPSS development. In 1997 HPSS received an R&D 100 Award from R&D Magazine for its scaleable architecture, its network-centric design which supports direct and network attached devices, and its parallel I/O capabilities.

HPSS was developed in response to the critical need for higher performance and larger capacities in data storage systems used in high-performance computing environments. With its emphasis on meeting the requirements for high end storage systems, HPSS is designed to store petabytes (10^{15}) of data and to transfer data at gigabytes (10^{9}) per second using network-connected storage devices.

As part of its network-centered design, HPSS provides servers and data movers that can be distributed across a high performance network to provide its 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 movers can transfer data in parallel streams for high aggregate bandwidth. For flexibility and performance, HPSS also allows for separate networks for control and data transfer.

HPSS runs on UNIX systems with no kernel modifications. The OSF Distributed Computing Environment (DCE) and TransArc's Encina are the basis for the distributed, transaction-based architecture. HPSS also uses the DCE Remote Procedure Call (RPC) mechanism for message control, the DCE Threads package for multitasking, and the DCE Security and Cell Directory Services. Along with Encina for transaction management and integrity, HPSS depends on the Encina Structured File Server (SFS) as the Metadata Manager.

HPSS supports storage hierarchies, which consist of multiple levels of storage. Files move up or down the hierarchy by migrate and stage operations based on usage patterns, storage availability, and other site policies. A storage hierarchy, or class of service, may consist of disk storage, followed by one or more levels of tape storage. Large files that are being archived may be written to a tape-only hierarchy.

The HPSS servers, shown in Figure 1, currently include the Name Server, Bitfile Server, Migration/Purge Server, Storage Server, Physical Volume Library and Repository, Mover, and Storage System Manager. In the next release, the Location Manager will allow the use of multiple geographically distributed Name Servers and HPSS systems using a common federated name space. As indicated in Figure 1, the servers may be replicated for improved performance. Details on each of the HPSS servers may be found in the references.

Figure 1 also shows the infrastructure components along the top. The HPSS user interfaces, or clients, are shown along the left side of Figure 1.

A number of user clients or interfaces are supported by HPSS. FTP and NFS version 2 are industry-standard interfaces. Parallel FTP (PFTP) supports standard FTP commands plus extensions to optimize performance by allowing data to be transferred in parallel. The HPSS Client API is a programming interface that allows programmers to take advantage of the specific capabilities of HPSS. HPSS can also act as an external file system to the IBM SP Parallel I/O File System. Additional interfaces available in the next release are for Distributed File System (DFS) and MPI/IO.

Configuration at Los Alamos

The High Performance Computing Environment at Los Alamos includes SGI/Crays and HPSS in both the secure and open networks. As shown in Table 1, the open SGI/Crays are configured as nodes with *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 SGI/Crays are configured as nodes with 64 MIPS R10K processors for a total of 1024 processors and 256 GB of memory. These configurations are periodically changed by splitting or merging nodes. The secure system will grow to 6144 processors with 1.5 TB of combined memory in 1999. The open system will grow to 1792 processors with 448 GB of memory.

As shown in Figure 3, each node in the open has two HIPPI connections to an internal switch fabric and one HIPPI connection for external network services, such as HPSS. Each node in the secure has four HIPPI connections to an internal switch and one HIPPI connection for external network services, such as HPSS data transfer. Additional external FIDDI connections are provided for full network services and HPSS control information.

A configuration diagram for HPSS is shown in Figure 2. The current equipment being deployed in the two networks is described in Tables 3 and 4. Capacity and usage information are also shown.

HPSS Performance

Our testing methodology is to determine the best possible performance between the clients and HPSS and between HPSS systems using ttcp tests and the best possible performance reading and writing to disk and tape. We then ran tests to determine the overhead associated with the HPSS infrastructure and finally full-blown HPSS tests.

Due to anomalies with protocol stack processing in AIX and Irix, we expected results to differ widely depending on buffer and window sizes. While we found many combinations that yielded bad performance, we also found a number of combinations that yield the performance given in Table 5. These window sizes are in the 32K-128K range, with comparable read and write sizes, and buffer sizes in the 256K-640K range, again with comparable read and write sizes. The best single stream result, 31 MB with the SGI/Cray as the sink and the mover as the source, was 20% higher than any of the other single stream results.

The device tests in Table 6 provide us with theoretically the best performance we should expect. Looking at the results from the two tables, we expect that a 43P mover will be able to support two IBM 3590 tape drives. Theoretically, the same mover can support a single SSA RAID subsytem for writes, but will be the limiting factor on reads.

Further tests were performed with HPSS client software to determine the overhead associated with the transaction and metadata management versus the actual data transfer component. These tests were run with the original IBM 580 HPSS server and the current IBM R24 server to see if there would be any performance improvement. The tests were also done using the server as the disk mover and using an IBM 43P as the disk mover. The results of these tests are in Table 7.

As expected, the infrastructure tests show that improvements in the HPSS server performance results in better transaction performance and increases the number of files that can be created or deleted. Transferring the data movement function to a separate system also improves transaction performance.

Additional tests were performed between the SGI/Cray client and HPSS. These tests used a locally-developed user interface called PSI. PSI runs on top of PFTP and would be expected to be slower than the previous tests. The results are shown in Table 8.

These results show that HPSS performance is much better with large files than small files. With small files, the speed of the transaction and data management 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 20% of the theoretical maximum. These tests were run using an SSA RAID adapter for the MCA bus. Enhanced adapters have been installed on PCI-based systems and subsequent tests should show an improvement in disk performance. Overall, much work remains to be done with buffer and socket sizes to determine the optimum sizes for maximum performance.

HPSS Requirements

The requirements for archival storage at Los Alamos are being driven by the ASCI Project. Historically, each year users have stored 750 times the available memory on the production computers. Also, users become frustrated if they are not able to store or retrieve their archived data within a reasonable period of time. We have estimated this performance requirement to be approximately one-half of the system's memory within a 20 minute time frame. These results are in Table 9.

Thus far the actual usage has been substantially below what has been estimated. The ASCI systems are more a capability than an actual production system. Utilization has been significantly below that expected from a production system. Also, our user community is just now learning how to use these systems as programs are being rewritten. These requirements probably could be moved out another year.

Summary

HPSS is currently in production in both our open and secure environments with usage growing at about 2TB/month overall. As HPSS has become a production system, testing has become more important to getting the optimum performance. Optimizing performance will be heavily dependent on managing buffer and socket sizes to maximize network performance. The next version of HPSS will include performance enhancements to the transaction and metadata management software. The deployment of new hardware at Los Alamos will further improve performance. Much work is still to be done to instrument the mover software and to develop a consistent test suite.

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. Available online at www.sdsc.edu/hpss. This paper contains many references to early HPSS work.

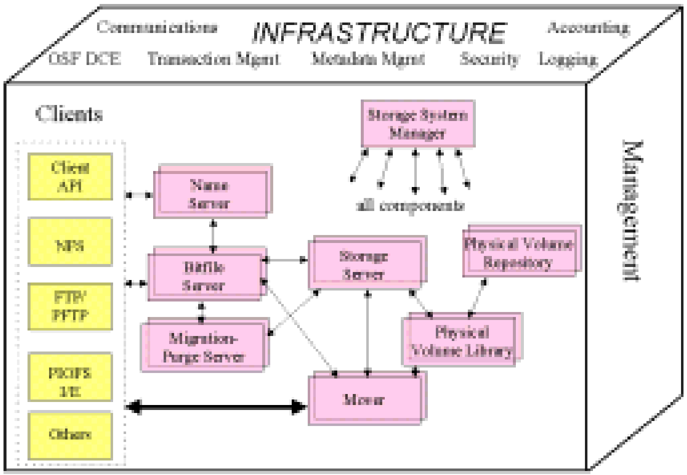

Figure 1. HPSS Infrastructure

Table 1. OPEN SGI/CRAY HIGH PERFORMANCE COMPUTING SYSTEM

Model	# CPUs	CPU Speed	CPU Cache	Memory
Origin 200*	2	180 MHz	1MB	256MB
Origin 200*	2	180 MHz	1MB	256MB
Onyx-2	64	195MHz	4MB	16GB
Origin 2000	32	195MHz	4MB	8GB
Origin 2000	32	195MHz	4MB	8GB
Origin 2000	32	195MHz	4MB	8GB
Origin 2000	96	195MHz	4MB	24GB
Origin 2000	128	195MHz	4MB	32GB
Origin 2000	128	195MHz	4MB	32GB
Origin 2000	128	195MHz	4MB	32GB
Origin 2000	128	195MHz	4MB	32GB

* Front-end system

Table 2. SECURE SGI/CRAY HIGH PERFORMANCE COMPUTING SYSTEM

Model	# CPUs	CPU Speed	CPU Cache	Memory
Origin 200*	2	180 MHz	1MB	256MB
Origin 200*	2	180 MHz	1MB	256MB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Origin 2000	64	195MHz	4MB	16GB
Onyx 2**	8	195MHz	4MB	4GB

* Front-end system ** Visualization server

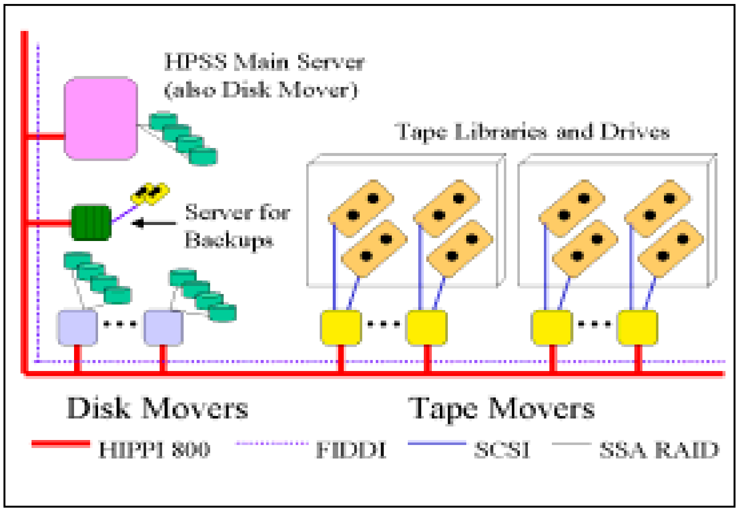

Figure 2. Los Alamos HPSS Configuration

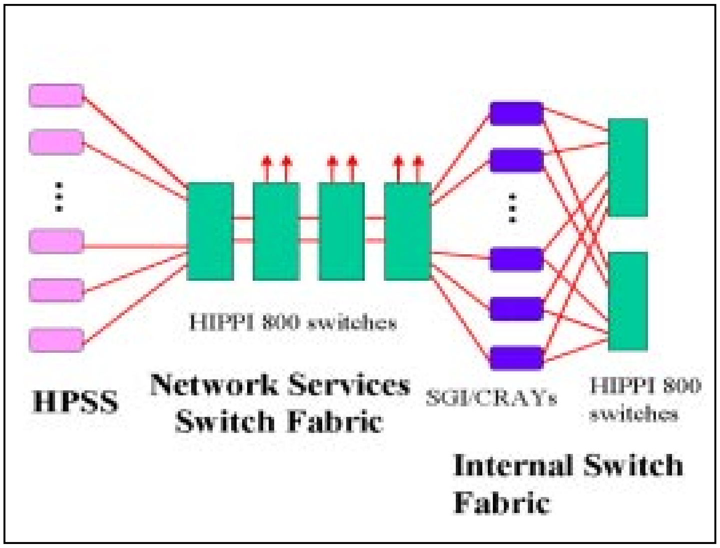

Figure 3. Los Alamos HIPPI 800 Network

Table 3.
Open HPSS Deployment and Usage Information

Description	Bus	Equipment	
Main Server (also disk mover)	MCA	1	IBM R24 w/ SSA RAID disks
Disk Movers	PCI	4	IBM 43P-240 w/ SSA RAID disks
Tape Movers	PCI MCA	9 1	IBM 43P-140 IBM R24
Tape Libraries		2	IBM 3494 (2 robots)
Tape Drives		16	IBM 3590
Server for MetaData Backups	PCI	1	IBM F30 w/ IBM 3570 tape library
Other Data			
Data Stored:	18 TB (1 TB/month growth)		
Files:	293 K (average size - 67 MB)		
Tape Capacity:	24 TB uncompressed 43 TB compressed (1.8 ratio assumed)		
Disk Capacity:	144 GB		

Table 4.
Secure HPSS Deployment and Usage Information

Description	Bus	Equipment	
Main Server (also disk mover)	MCA	1	IBM 580 w/ SSA RAID disks
Disk Movers	PCI	2	IBM F30 w/ SSA RAID disks
Tape Movers	MCA	2	IBM R24
Tape Libraries		2	IBM 3494 (2 robots) STK 4400 (shared, 3 robots)
Tape Drives		7 5	IBM 3590 STK Timberline
Server for MetaData Backups	PCI	1	IBM F30 w/ IBM 3570 tape library
Other Data			
Data Stored:	6 TB (1 TB/month growth)		
Files:	139 K (average size - 38 MB)		
Tape Capacity:	17 TB uncompressed 31 TB compressed (1.8 ratio assumed)		
Disk Capacity:	180 GB		

Table 5. TTCP (memory to memory) Tests

Description	Result (single stream)	Result (dual stream)
Between SGI/Cray nodes	71 MB/sec	NA
Write to 43P mover	20 MB/sec	24MB/sec
Read from 43P mover	31 MB/sec	27 MB/sec
Write to 43P mover (2 cpu)	NA	32MB/sec
Read from 43P mover (2 cpu)	NA	32 MB/sec
Note: Other system was SGI/Cray or Cray M98		

Table 6. Device Tests

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

Table 7. 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 / sec	.75-.90 files / sec	2.3-3.4 files / sec
IBM R24 Server (disk mover on server)	4.6-5.8 files / sec	1.6-1.9 files / sec	5.2-5.8 files / sec
IBM R24 Server (IBM 43P disk mover)	5.0-5.8 files / sec	2-2.4 files / sec	6.0-6.6 files / sec

Table 8. Client to HPSS Tests

Optn	Source	Sink	File size	Result
Write	client mem	HPSS tape	1GB	11.40 MB/s
Read	HPSS tape	client mem	1GB	11.59 MB/s
Write	client disk	HPSS tape	1GB	11.00 MB/s
Read	HPSS tape	client disk	1GB	11.65 MB/s
Write	client disk	HPSS disk	500MB	6.60 MB/s
Read	HPSS disk	client disk	500MB	16.37 MB/s
Write	client disk	HPSS disk	50MB	5.60 MB/s
Read	HPSS disk	client disk	50MB	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)

Table 9. ASCI Data Storage Requirements

	1998	1999	2000	2001	2002	2003	2004
ASCI memory (TB)	.444	1.5	4.5	8.9	15.8	28	50
Storage Growth / Year (PB)	.3	1.1	3.4	6.7	12	21	37.5
Total Storage Capacity (PB)	5.6	1.7	5	12	24	45	82
Single File Xfr Rate (GB/sec)	0.2	0.6	2	3.7	6.6	12	21