High Performance Storage System at Sandia National Labs

R. Michael Cahoon, Scientific Computing Systems, Sandia National Laboratories

ABSTRACT: *Scientific computing centers are acquiring large, distributed memory machines. With memory systems of .25 to 2.5 terabytes, these machines will deliver 1-10 teraflop computing capabilities. The need to move 10's or 100's of gigabytes, and the need to provide petabyte storage systems are issues that must be addressed before the year 2000. Work currently underway at Sandia addresses these issues. The High Performance Storage System (HPSS) is in limited production and the mass storage environment to support Sandia's teraflop computer system is being constructed.*

1 Introduction

Sandia National Laboratories, a US Department of Energy laboratory, works with the US Defense Department, the National Science Foundation, NASA, and industry to develop High Performance Computing (HPC) technologies and apply them to nationally important problems. Its mission includes national security, industrial competitiveness, energy resources, and environmental quality. At Sandia, computers are being used to design and optimize materials ranging from catalysts to opto-electronics, and simulations are replacing tests and experiments that are environmentally unacceptable or prohibitively expensive.

Figure 1. High Performance Storage System

Sandia uses a variety of computing resources to address the wide-range of computational problems investigated at the laboratories. The acquisition of Sandia's Teraflop Compute Server and rapidly enlarging data sets will require unprecedented levels of data storage and access. Today, data sets of hundreds of gigabytes (GB) are not unusual, nor are total system storage requirements of 10 or more terabytes (TB). The future demands data sets of TB size and total storage requirements of petabytes (PB). Sandia is addressing this requirement thru the development of a scalable, network-centered, parallel storage system. Figure 1 shows the initial system being deployed. The mission of the Scientific Computing Systems department is to provide such a capability and ensure its integration into the Sandia environment.

2 High Performance Storage System (HPSS)

The rapid growth in the size of datasets has caused a serious imbalance in I/O and storage system performance and functionality relative to application requirements and the capabilities of other system components. The High-Performance Storage System (HPSS) is a scalable, next-generation storage system that will meet the functionality and performance requirements of large-scale scientific and commercial computing environments.

2.1 HPSS Partnership

To achieve this task of providing scalable storage, Sandia, in partnership with other DOE labs, NASA laboratories, and IBM is developing software systems to manage the next generation of mass-storage technology. No one organization has the will or resources to under take such a large software development program. Primary development responsibility resides with four U. S. Department of Energy National Laboratories (Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Sandia National Labora-

tories) and IBM Government Systems. Other developers include NASA-Langley Research Center, NASA-Lewis Research Center, and Cornell University. In addition to the DOE Labs and NASA centers, early deployment partners each have very challenging heterogeneous environments and research and development agendas that add significant value to HPSS. The sites include: Argonne National Laboratory, California Institute of Technology teamed with the Jet Propulsion Laboratory, Cornell Theory Center, Fermi National Laboratory, Maui High Performance Computer Center, San Diego Supercomputer Center and the University of Washington.

2.2 HPSS Design

Detailed technical descriptions of the design are contained in references [2],[3]. An overview, extracted from these references will be presented here. The network-centered HPSS architecture, based on the IEEE Mass Storage Reference Model version 5 ([3]), has a high-speed network for data transfer and a logically separate network for control (Figure 2)([4],[5],[6],[7], [8]). The control network uses the Open Software Foundation's (OSF) Distributed Computing Environment (DCE) Remote Procedure Call technology [9]. In actual implementation, the control and data-transfer networks may be physically separate or shared.

An important feature of HPSS is its support for both parallel and sequential I/O and standard interfaces for communication between processors (parallel or otherwise) and storage peripherals. In typical use, clients direct a request for data to an HPSS

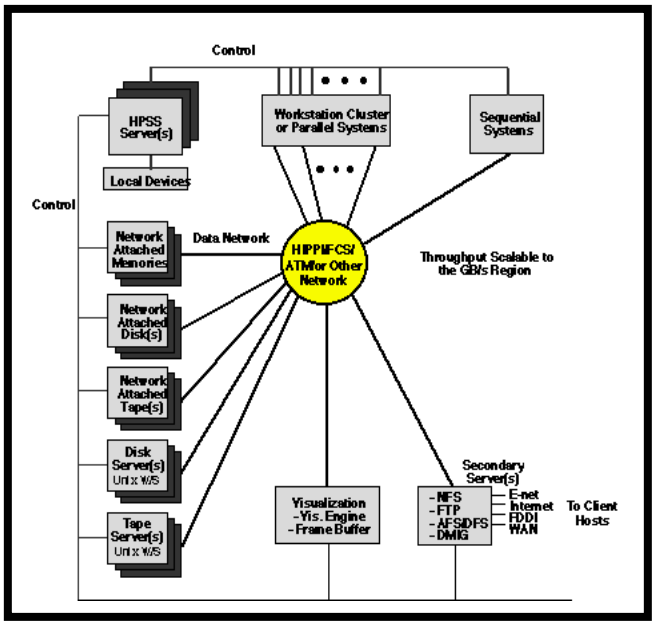

Figure 2. Typical Configuration

server. The server directs network-attached storage peripherals or servers to transfer data directly, sequentially or in parallel, to the client node(s) through the high-speed, data-transfer network (Figure 2). HPSS also supports devices attached to HPSS servers. TCP/IP sockets and IPI-3 over a High-Performance Parallel Interface (HIPPI) are being utilized today; Fibre Channel Standard (FCS) with IPI-3 or SCSI, or Asynchronous Transfer Mode (ATM) will be supported in the future ([10],[11],[12],[13],[14]). Through its parallel storage and I/O support by data striping, HPSS will continue to scale upward as additional storage peripherals and controllers and network connectivity are added.

The HPSS components (Figure 3) are multithreaded using DCE threads [9] and can be distributed and multiprocessed. Multithreading is also important for serving large numbers of concurrent users. HPSS also uses DCE security and distributed time and directory services. HPSS uses the Transarc Encina software for support of atomic transactions, logging, and system metadata ([12],[15],[16]). The storage peripherals managed by HPSS can be organized into multiple storage hierarchies ([4],[17]). Storage-system management (SSM), built around an ISO-managed object framework, is another important HPSS focus ([18],[19],[20]). HPSS components can be run either on single or distributed server machines or on one or more nodes of a parallel system.

2.3 Key Features

The key objectives of HPSS are:

- Scalability in several dimensions, including distribution and multiprocessing of servers, data transfer rates to gigabytes per second, storage capacity to petabytes, file sizes to terabytes, number of naming directories to millions, and hundreds to thousands of simultaneous clients.
- Modularity by building on the IEEE Reference Model architecture (Figure 3) to support client access to all major system subcomponents, replacement of software components during the storage system's life cycle, and integration of multivendor hardware and software storage components.
- Portability to many vendors' platforms by building on industry standards, such as the OSF DCE, standard communications protocols, C, POSIX, and UNIX with no kernel

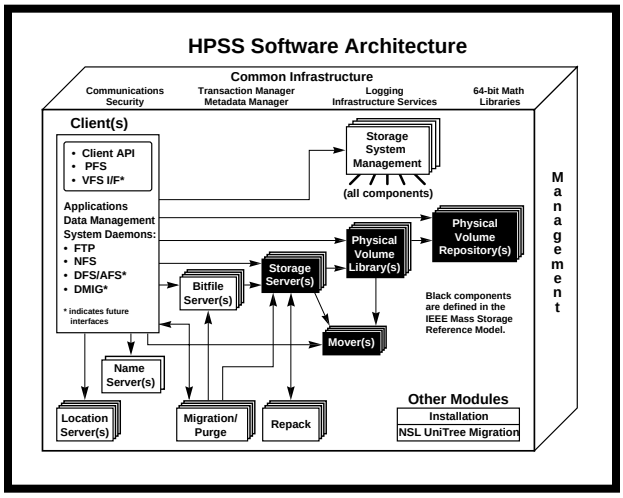

Figure 3. HPSS Software Model Diagram

modifications. HPSS uses commercial products for system infrastructure, and all HPSS component interfaces have been put in the public domain through the IEEE Storage System Standards Working Group.

- * Reliability and recoverability through support for atomic transactions among distributed components, mirroring and logging of system metadata or user data, recovery from failed devices or media, reconnection logic, ability to relocate distributed components, and use of software engineering development practices.
- * Client APIs to all major system components, including a parallel Client API and interface to vendor parallel file systems, and support for industry-standard services such as File Transfer Protocol (FTP) (sequential and parallel) and Network File System (NFS) ([11],[21]). Future support will include OSF's Distributed File System (DFS), and the Unix Virtual File System (VFS) [22]. Support is also planned for interface to local and distributed file systems through the Data Management Interface Group (DMIG) standard ([23],[25]).
- * Support for better integration with data-management systems through appropriate interface functionality at multiple levels in the architecture.
- * Security through DCE and POSIX security mechanisms, including authentication, access control lists, file permissions, and security labels.
- * System manageability through a managed-object reporting, monitoring, and database framework, and management operations with graphical user interface (GUI) access and control.
- * Distributability by building on a client/server architecture and use of an OSF DCE infrastructure.

2.4 HPSS Software Model

The HPSS software model is shown in Figure 3 and fully described in references [1],[2]. Primary storage is a file of logical sequence of 2^{64} bytes. In HPSS terms this is a bitfile. A user can read, write, or seek to any part of a bit file by first contacting the name server. The Bitfile server supports both sequential and parallel read/writes of data to bitfiles. In conjunction with Storage Servers, the Bitfile Server maps logical portion of bitfiles onto physical storage devices. The Storage Server provides a hierarchy of storage objects: logical storage segments, virtual volumes and physical volumes. The Storage Server(s) in conjunction with the Mover(s) have the main responsibility for HPSS's parallel I/O operations. The Mover is responsible for transferring data from the source device(s) to sink device(s). Due to the inherent scalability of HPSS there can be multiple servers and movers.

3 Status of Development Effort

Release 1, tape only, is currently available and in use or integrated into production environments at several sites. Release 3 (there will be no release 2) is scheduled for general availability during the second quarter of 1996. Specifications for Release 4 are near completion with the release scheduled for second quarter 1997. Release 4 will focus on scalability and performance issues.

From a software development perspective, HPSS is a large project. The project is even more unique in that is a collaboration between many organizations. Very few face-to-face meetings are scheduled. Most of the coordination is done via email or weekly teleconference calls. Regular reporting and testing programs have been established and the results shared via email. Figure 4 shows the various modules of HPSS. The development responsibility for a module is assigned to development organization. In addition to their own module responsibilities, IBM provides the rigor of a commercial software development project.

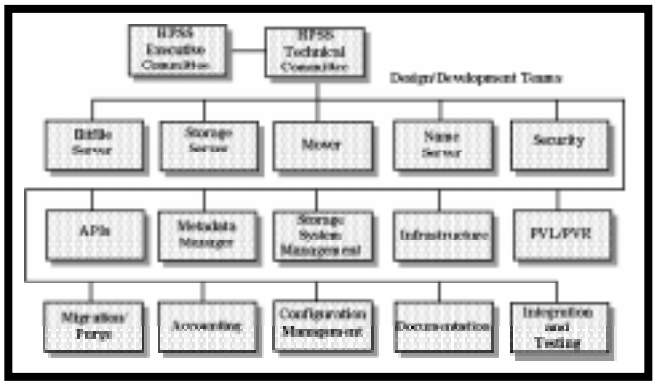

Figure 4. HPSS Organization

4 HPSS Deployment Sites

Early operational experiences for a wide range of environments and platforms has contributed significantly to the development effort. These early adopters have provided valuable feedback as to the reliability and usability of HPSS. The current plans and status of the various deployment site are reviewed [26]:

Sandia National Laboratories will move aggressively to use HPSS as the primary storage software for its 1840-node Intel Paragon. Noteworthy is the SNL decision to use ATM connectivity to move files into both tape (IBM 3494 Tape Library with eight 3590 tape drives) and disks (IBM 7135 disk array). Sandia is deploying an OC12 interface in the Intel Paragon and connecting to the HPSS system via a FORE systems switch. Additional ATM connectivity to FDDI networks and other computing platforms will also be provided.

Lawrence Livermore National Laboratory includes several facilities in which HPSS will have a major role.

Scalable I/O Facility (SIOF)

The SIOF provides an environment for testing I/O systems and strategies. The software architecture includes a parallel I/O API based on MPI-IO layered over HPSS. FCS provides parallel

connectivity from a Meiko CS-2 to an array of tape and disk controllers.

Livermore Computing (LC)

LC handles the production computing for LLNL. HIPPI and FCS will provide parallel storage connectivity for the Meiko into network-attached arrays of both disks and tape. In addition, storage requirements of workstation clusters and other LANS will be served by HPSS.

National Energy Research Supercomputer Center (NERSC)

NERSC has a large user group distributed across the country. Projected storage requirements reach the PB range in 1997. Storage software currently includes the Common File System (CFS) and NSL-Unitree. Migration to HPSS is projected for 1997. The current hardware configuration includes tape and disk arrays connected through HYPERchannel for CFS and HIPPI for NSL-Unitree. Cray computers, including a C-90, provide the computing power. The transition to HPSS will include connectivity provided through HIPPI or FCS. NERSC is also projecting the addition of a powerful MPP machine on the same time scale to complement the C-90.

Cornell Theory Centered (CTC) is moving aggressively to bring HPSS into production or near-production status. The principal file source will be the 512-node IBM SP2, with 80 GB of memory, 1.2 TB of disk, and rated peak computing capacity of 136 gigaFLOPS. Network connectivity will include all of HIPPI, ATM, HPS, and FIDDI. The AFS shared file system and PIOFS parallel file system are used. CTC is projecting storage system migration from NSL-Unitree to HPSS in 1996, and will also pursue test-bed activities, an example being studies of AIMNET.

Maui High-Performance Computing Center (MHPCC) is another major center featuring IBM equipment, the MHPCC SP2 includes 400 nodes, 56 GB of memory, 784 GB of disk, and an additional 125 GB of NFS storage. A dedicated Essential HIPPI switch is available for HPSS testing and 10 HIPPI nodes are incorporated into the SP2. The hardware includes IBM 3490 E and NTP tape drives and Maximum Strategy GEN-5 disks. NSL-Unitree is the current production software, with HPSS testing now under way.

Los Alamos National Laboratory (LANL) is currently running an HPSS development system with both disk and tape, and will have a production system in 1996. The initial storage hardware will be four STK Timberline drives, four IBM 3590 drives in a 3494 library, and a disk array, serving an SMP-type machine, several Crays, a workstation cluster, and individual workstations. Networks will include FDDI, ethernet, and HiPPI.

Center for Computing Sciences (CCS) at ORNL computing environment includes a spectrum of machines (Intel Paragons: 66 GP-node XP/S 5, 512 GP-node XP/S 35, and 1024 MP3-node XP/S 150; a 16-node IBM SP2; and a KSR1-64) and a storage complement including a 100 TB IBM 3495 tape robot, soon to have eight NTP drives, 216 GB of Maximum Strategy disks, a 14.2 TB Storage Tek Silo (four Timberline and 2 Redwood drives with Powerhorn robotics), and a CREO optical tape system. While awaiting a fully configured ESCON-attached 3495 library, CCS is using a SCSI-attached 3494 library with four NTP drives for HPSS testing. The current production storage system is NSL-Unitree. At this point, HIPPI and ESCON switches provide network connectivity.

Storage systems, features, and development are a principal focus in the CCS, examples being the creation of the UTI interface for NSL-Unitree and the expansive storage system just described. Accordingly, a test-bed environment, separate from the production environment, has been established in which HPSS can be tested, even stressed. Tests of HPSS Release 1 utilized the Intel Paragon XP/S 5 and HIPPI/ESCON connections to both IBM and Storage Tek libraries, the latter configured with six parallel paths. For Release 3 tests, the system will be extended to include four parallel HIPPI paths into the Maximum Strategy RAID array and eight ESCON paths into the IBM 3495 library.

5 Sandia Deployment Efforts

Sandia's initial deployment system, in support of the Intel Paragon, is described above. Future plans call for the deployment of classified and unclassified systems to support the Intel Teraflop system. Since the operational mode of the Teraflop system will be to run a single job, initial criteria requires that 150 gigabytes of main memory be moved to HPSS in a desired time of 15 minutes.

Sandia's next generation system will utilized low-cost workstations as Mover platforms, replacing the IBM 590/570s currently used. It is anticipated that these mover platforms will be PCI based with ATM interfaces. Multiple tape drives connected to multiple Movers will interface to OC12, eventually OC48, interfaces on the Teraflop compute server. Current OC3 experience indicates that a memory to memory copy on IBM 590/570s acheive 120/80 Mb/s respectively.

Since Sandia requires HPSS to meet the I/O requirements of its Intel Paragon, Sandia worked early in the development cycle with ORNL to develop a parallel FTP. Starting with the public domain standard FTP client and daemon code, the code was modified to support multiple Movers for handling parallel data transfers and extended commands [2].

Earlier experience at Sandia indicates that using a one-stripe PFTP is faster than using native FTP. Current testing results indicate that HPSS will be able to achieve the performance criteria described above at an affordable cost. Test results are shown in Figure 5. These test establish a base-line for future development and deployment efforts. The results are quite positive and will guide future directions and decisions.

6 More Information

Extensive information about HPSS is available over the World Wide Web. Connect to the web site http://www.ccs.ornl.gov/hpss. For more information about HPSS, contact Dick Watson, LLNL, 510-422-9216, dwatson@llnl.gov; or Bob Coyne, IBM Government Systems, 713-335-4040, coyne@vnet.ibm.com

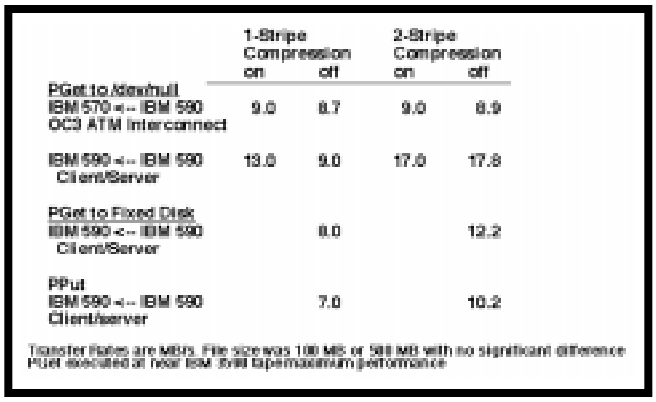

Figure 5. IBM 3590 Performance

7 Acknowledgments

The HPSS project is a very large software project with over 30 active participants at numerous, geographically dispersed sites. Any technical information contained in this report is their work and details can be found in the appropriate references. Thanks to the Sandia team, Rena Haynes, Bill Rahe, Marty Barnaby, Hilliary Jones, Jerry Bollig, Louie Martinez, and Jim Laros for making HPSS a reality at Sandia. This work was supported by the Department of Energy under Contract DE-AC04-94AL85000.

8 References

[1] D. Teaff, R. Coyne, and R. Watson, "The Architecture of the High Performance Storage System," 4th NASA GSFC Conf. Mass Storage Systems and Technologies, College Park, MD, Mar. 28-30, 1995.

[2] R.W. Watson and R.A. Coyne, “The Parallel I/O Architecture of the High Performance Storage System (HPSS),” Proc. Fourteenth IEEE Symposium on Mass Storage Systems, Monterey, CA, September 1995, pp. 27-44

[3] IEEE Storage System Standards Working Group (SSSWG) (Project 1244), "Reference Model for Open Storage Systems Interconnection, Mass Storage Reference Model Version 5," Sept. 1994. Available from the IEEE SSSWG Technical Editor Richard Garrison, Martin Marietta (215) 532-6746.

[4] R.A. Coyne, H. Hulen, and R.W. Watson, "Storage Systems for National Information Assets," Proc. Supercomputing 92, Minneapolis, MN, Nov. 1992, pp. 626-633.

[6] B. Collins et al., "Los Alamos HPDS: High-Speed Data Transfer," Proc. 12th IEEE Symp. Mass Storage Systems, Monterey, CA, Apr. 1993.

[7] R. Hyer, R. Ruef, and R.W. Watson, "High Performance Direct Network Data Transfers at the National Storage Laboratory," Proc. 12th IEEE Symp. Mass Storage, Monterey, CA, IEEE Computer Society Press, Apr. 1993.

[8] R.H. Katz, "High Performance Network and Channel-Based Storage," Proc. IEEE, Vol. 80, No. 8, pp. 1238-1262, Aug. 1992.

[9] M. Nelson et al., "The National Center for Atmospheric Research Mass Storage System," Digest of Papers, 8th IEEE Symp. Mass Storage Systems, May 1987, pp. 12-20.

[10] Open Software Foundation, Distributed Computing Environment Version 1.0 Documentation Set. Open Software Foundation, Cambridge, MA, 1992.

[11] G.S. Christensen, W.R. Franta, and W.A. Petersen, "Future Directions of High-speed Networks for Distributed Storage Environments," Digest of Papers, 11th IEEE Symp. Mass Storage Systems, Oct. 7-10, 1991, IEEE Computer Society Press, pp. 145-148.

[12] Internet Standards. The official Internet standards are defined by RFC's (TCP protocol suite). RFC 783; TCP standard defined. RFC 959; FTP protocol standard. RFC 1068; FTP use in third-party transfers. RFC 1094; NFS standard defined. RFC 1057; RPC standard defined.

[13] T.W. Tyler and D.S. Fisher, "Using Distributed OLTP Technology in a High Performance Storage System," Proc. 14th IEEE Symp. Mass Storage Systems, Monterey, CA, Sept. 1995.

[14] L.D. Witte, "Computer Networks and Distributed Systems," IEEE Computer, Vol. 24, No. 9, Sept. 1991, pp. 67-77.

[15] D.E. Tolmie, "Local Area Gigabit Networking," Digest of Papers, 11th IEEE Symp. Mass Storage Systems, Oct. 7-10, 1991, IEEE Computer Society Press, pp. 11-16.

[16] S. Dietzen, Transarc Corporation, ""Distributed Transaction Processing with Encina and the OSF/DCE," Sept. 1992, 22 pages.

[17] B.W. Lampson, "Atomic Transactions," in Distributed Systems Architecture and Implementation, Berlin and New York: Springer-Verlag, 1981.

[18] A.L. Buck and R.A. Coyne, Jr., "Dynamic Hierarchies and Optimization in Distributed Storage System," Digest of Papers, 11th IEEE Symp. Mass Storage Systems, Oct. 7-10, 1991, IEEE Computer Society Press, pp. 85-91.

[19]"Information Technology-Open Systems Interconnection-Structure of Management Information-Part 4: Guidelines for the Definition of Management Objects," ISO/IEC 10165-4, 1991.

[20] ISO/IEC DIS 10040 Information Processing Systems-Open Systems Interconnection-Systems Management Overview, 1991.

[21] S. Louis and R. Burris, "Management Issues for High Performance Storage Systems," Proc. 14th IEEE Computer Society Mass Storage Systems Symp., Sept. 1995.

[22] R. Sandberg et al., "Design and Implementation of the SUN Network File System," Proc. USENIX Summer Conf., June 1989, pp. 119-130.

[23] S. Kleiman, "V nodes: An Architecture for Multiple File System Types in SUN UNIX," Proc. Summer USENIX Conf., 1986, Atlanta, GA.

[24] Data Management Interfaces Group, "Interface Specification," Version 2.0, Nov. 1994.

[25] P. Lawthers, "Data Management Applications Programming Interface," Proc. 14th IEEE Mass Storage Symp., Sept. 1995.

[26] K. L. Kliewer, private communications