SANDIA REPORT

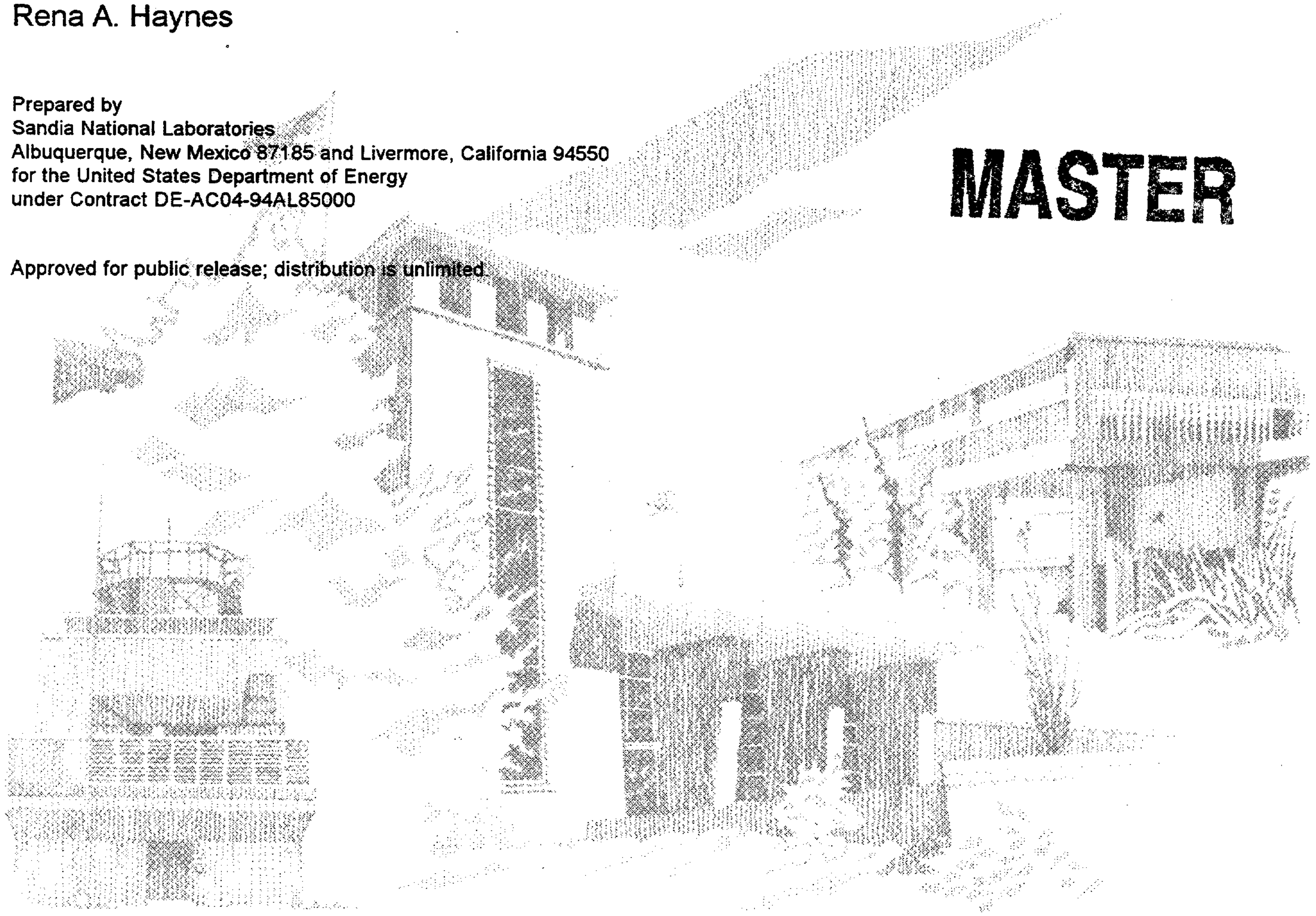
SAND96–1724 • UC–705
Unlimited Release
Printed January 1997

Hierarchical High-Performance Storage System Testbed Project at Sandia National Laboratories

Rena A. Haynes

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550
for the United States Department of Energy
under Contract DE-AC04-94AL85000

Approved for public release; distribution is unlimited.



Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

SAND96-1724
Unlimited Release
Printed January 1997

Distribution
Category UC-705

Hierarchical High-Performance Storage System Testbed Project at Sandia National Laboratories

Rena A. Haynes

Scientific Computing Systems
Sandia National Laboratories
Albuquerque, New Mexico 87185-0807

Abstract

The Hierarchical High-Performance Storage System (HPSS) Testbed project at Sandia National Laboratories was part of a research collaboration between industry, national research centers, and national laboratories to develop mass storage system software that would scale to meet the capacity and performance required by supercomputer and massively parallel computational environments. This report describes the software that was developed within this collaboration as a result of a cooperative research and development agreement between Sandia National Laboratories and International Business Machines (IBM) Corporation, Government Systems.

Acknowledgement

The work described in this report was accomplished due to the efforts of software development teams at IBM and Sandia. Special thanks to Danny Teaff, Kurt Everson, and Rich Ruef on the IBM team for their development of the basic logging, native interface library, and parallel file transfer components of HPSS. Thanks to the Sandia team of Bill Rahe, Marty Barnaby, Hilary Jones, and Sue Kelly for their development and extensive software testing efforts.

Table of Contents

Introduction

The primary objective of the Hierarchical High-Performance Storage System Testbed project was to develop a High Performance Storage System (HPSS) that would be of general use in a computer network environment ranging from massively parallel processor (MPP) systems to desktop workstations. Since the project inception, improvements in application algorithms, processing power, memory sizes, data collection devices, multimedia capabilities, and integration of enterprise data at Department of Energy national laboratories have produced the need for storage systems with three orders of magnitude improvements in capacity and performance.

As computer technology continues to advance, storage requirements increase proportionately. An analysis of storage requirement trends for scientific computation suggests that the amount of data being stored on-line is increasing by 100 percent annually [1]. To address the constant push on storage systems capacity and performance, the project proposed the development and demonstration of scalable mass storage system software. The key idea was to provide a mass storage system scalable in terms of performance and capacity by implementing modular software components that could have duplicate instances or be replaced as new technology is introduced. With this architecture, input/output performance can be increased through the use of multiple parallel transfer paths.

Cooperative Research and Development Agreement (CRADA) No. SC93/1198 was part of a unified collaboration between national laboratories— Sandia (SNL), Los Alamos (LANL), Lawrence Livermore (LLNL), and Oak Ridge (ORNL); research centers— Cornell University and NASA Langley Research Center; and International Business Machines Corporation (IBM) Government Systems. While each of the laboratories and research centers brought unique mass storage systems, supercomputing experience, and hardware to the collaboration, Sandia provided expertise in developing mass storage systems security, managing mass storage systems in a large network environment, and developing massively parallel algorithms. IBM provided project coordination, software integration facilities, and commercialization support as well as the development of key HPSS components. Sandia provided software development, including design, code development, code testing and system testing.

Description

The HPSS software that was developed as a result of the unified collaboration included development of twenty major software components as well as infrastructure and interface software. The HPSS architecture is described in [2] and [3]. This report describes the software that was developed as a result of CRADA SC93/1198. The software was developed for five task areas— HPSS communications infrastructure, HPSS Name Server, HPSS security, HPSS logging, and HPSS client interfaces. Software developed for these tasks was demonstrated at the

Association for Computing Machinery (ACM)/ Institute of Electrical and Electronics Engineers (IEEE) Supercomputing Conferences in 1994 and 1995.

HPSS Communication Infrastructure

HPSS communication infrastructure software provides the capabilities needed for the distributed components of HPSS to communicate and function in a reliable manner. While the Open Software Foundation Distributed Computing Environment (DCE) protocol [4] was selected as the mechanism used for HPSS control communications, software was developed to enable maintenance of logical sessions, or connections, between HPSS server components.

The communication infrastructure software package is used by all DCE-based HPSS components. An HPSS server that needs to communicate with another will open a connection to that server. The connection management software in both servers will periodically communicate to determine whether the other server is available. If one of the servers terminates, the peer server is notified so that it may clean up any shared state.

An additional software component was developed under this task to enable startup, shutdown, and termination detection of HPSS server components. This component, called the startup daemon, runs on every computer where an HPSS server runs. The startup daemon shares responsibility with each HPSS server for ensuring that only one copy of the server is running at a given time. This helps the HPSS system administrative component determine whether servers are still running, and allows the HPSS administrator to send signals directly to the servers.

HPSS Name Server

A major component of HPSS is the server that provides a human understandable name and a logical directory tree structure for data sets contained within an HPSS installation. This component, called the name server, is also responsible for maintaining ownership and access permissions to HPSS data sets. Software was developed to implement the name server component for the first release (R1) of HPSS, which was demonstrated at Supercomputing 1994.

While the R1 HPSS name server provided adequate functionality, additional scalability was required. Sandia participated in a follow-on task to develop a fully scalable HPSS name server.

HPSS Security

HPSS software security provides mechanisms that allow HPSS components to communicate in an authenticated manner, to authorize access to HPSS objects, and to enforce access controls on HPSS objects. Software that maintains security contexts between HPSS client and server components was developed. Information contained in security contexts include identity, location, and authorization information about the client. Additional identity information about the user originating the request is also available. Security contexts also enable maintenance of a session key that can be used to protect or certify the integrity of data passed between the client and server.

Security context information is obtained from the DCE authenticated remote procedure calls (RPC) and security registry records. This information along with DCE access control entries are used to protect access to HPSS server interfaces.

Protection of access to HPSS data sets is provided through a distributed mechanism whereby a requester's access permissions to an HPSS file object is specified by the HPSS data set authorization agent, the name server. These permissions are processed by the HPSS data set authorization enforcement agent, the bitfile server. The integrity of the access permissions is certified by the inclusion of a checksum which is encrypted using a security context key shared between the HPSS name server and bitfile server components.

An additional security mechanism was provided to allow identification of end users through a site supplied authentication policy.Site policy is implemented though an HPSS policy manager component described in the next paragraph. A sample site identification mechanism using an interface to DCE security services was developed. The HPSS file transfer interface components based on the File Transfer Protocol (FTP) [5] were augmented to enable utilization of this mechanism.

The basic architecture of the policy manager is to provide DCE RPC interfaces to routines that input opaque data. These HPSS supplied routines call your site supplied routines to process the input data and generate reply data that is returned through the RPC interface to the calling HPSS server. HPSS operations that support site policy will call the appropriate HPSS policy manager application programming interface (API) and alter its operation based on the reply information.

HPSS Logging

HPSS logging capabilities provide a common mechanism for HPSS components to record access, error, and security event information that can be retrieved and processed at a later date. Software to implement the basic logging capabilities was developed. Security auditing software which utilizes the basic logging software was also developed. The security audit capability formats security event records that contain information about the type of security event, the security event subject, and the security event object. Software to implement a mechanism to enable or disable generation of each type of security event based on site policy was also developed.

Security events that may be audited at each server include authentication, file object creation, deletion, permissions changes, and file data access. HPSS servers can be configured to produce an audit record whenever one of these events occurs, to produce an audit record whenever one of these events occurs with a failed result, or to bypass auditing a security event.

HPSS Client Interfaces

HPSS client interface software provides computer users with mechanisms to store, retrieve, and manage data sets in an HPSS system. HPSS client interfaces are based on standards developed in the computing community. A core library of functions was developed to implement a native HPSS interface compatible with the Portable Operating System Interface (POSIX) standard [6].

To allow access from users on systems unable to utilize the native HPSS interface, software was developed to implement file transfer protocol (FTP) and Network File System (NFS) interfaces. The FTP interface provides a universal mechanism for storing and retrieving data to and from HPSS. This protocol was augmented to allow parallel data transfer as well as additional data set management capabilities. Parallel client FTP software was implemented for IBM, Intel, Cray, Silicon Graphics, Sun Microsystems, and Meiko computer platforms.

NFS Version 2 is another widely used mechanism to manage and access data [7]. NFS provides functionally transparent access to HPSS file objects. Software was developed that implements the NFS protocol, manages attribute and data caches, and collects and exports operational and caching statistics. In addition, software was developed to implement a user mapping mechanism whereby access to HPSS file objects can be controlled by requiring a user to first obtain an entry in the NFS map cache. A utility to insert, remove, and list NFS map cache entries was also developed.

Benefits to DOE

Benefits derived from this project can only be discussed within the framework of the unified collaboration. For DOE, the unique characteristics and contributions of the unified project have:

- Yielded early technology transfer using commercial products;
- Benefited Sandia, LLNL, LANL, and ORNL, and other DOE and government collaborating users because these users have set requirements and priorities;
- Established a close connection between the participants in the project and the IEEE storage system standards work;
- Created a unified, distributed, hierarchical storage system supporting multiple storage hierarchies;
- Recognized the importance of developing distributed, standards-based storage system management tools;
- Benefited from collaborating with multiple DOE and other supercomputer sites.

The significant benefits to DOE and Defense Programs (DP) are as follows: scalable data transfer support to achieve orders of magnitude improvement in performance, explicit support for massively parallel computing, more modular system design, wider use of standards, and increased support for scientific data management. All of the DOE laboratories in the collaboration, including the DP laboratories, are currently deploying HPSS. Having a high-performance storage system such as HPSS available to massively parallel and large memory systems at these facilities is critical to maintaining balanced input/output performance, thus getting the full productivity from the hardware and user communities.

HPSS has been selected as the archival component for the Accelerated Strategic Computing Initiative (ASCI), which is an integrating element of the Stockpile Stewardship Program to shift from nuclear test-based methods to computational-based methods.

As a result of this project, improved and state-of-the-art systems architecture for high performance storage systems has been developed. The HPSS collaboration has grown to include eleven sites that are currently deploying HPSS software.

Economic Impact

With respect to commercialization of the results, IBM plans to implement eleven early deployment sites among universities, federal programs, and supercomputer centers. HPSS early deployment is planned for third and fourth quarters of 1996. Early deployment will include beta testing and early production support (i.e., pilot programs) or proof-of-concept demonstrations. After successful early deployment, HPSS is planned for commercial release in the first quarter of 1997.

The early deployment sites are:

Caltech/Jet Propulsion Laboratory
Cornell Theory Center
Fermi National Accelerator Laboratory
Lawrence Livermore National Laboratory
Los Alamos National Laboratory
Maui High Performance Computing Center
Oak Ridge National Laboratory
San Diego Supercomputer Center
Sandia National Laboratories
NASA Langley Research Center
University of Washington

HPSS addresses the federal and commercial customer requirement for higher performance storage system software. HPSS software would not have been developed and deployed without the cooperation of the development partners and early deployment sites. The expense and skill required to provide a commercial product with HPSS features and functions would not have been a justifiable investment given the supercomputing market segment maturity. The industry is better off because the HPSS project has produced a software service that would not have otherwise been developed at this point in the market evolution but which is needed and which will support growth of the entire supercomputing market segment.

One of the problems to be overcome is the acceptance and support of open systems by the High Performance Computing and Communication (HPCC) vendor community. HPSS is based largely on DCE infrastructure as well as the Encina transaction processing software from Transarc Corporation. The availability of these software packages is required to port HPSS servers to more high-end computing platforms. Until these standards are more firmly embraced by the HPCC vendor community, HPSS server implementation would be limited to about four vendor operating system platforms.

Another problem is providing test-beds large enough to test the limits of HPSS scalability. Sites with enough resources and capacity to test systems larger than one petabyte are scarce and heavily used. It is not practical to make these facilities available to the HPSS project. Alternatively, HPSS scalability will continue to be tested at the very high end as customers evolve the HPSS supported systems. This approach will require multiple years of field support before HPSS has completed scalability testing.

The principal projected advantage is the availability of a scalable storage system that provides single file and aggregate file transfer rates in the range of multiple billions of bytes per second. This advantage is being tested during the remainder of 1996.

A commercially-available, high-performance storage system is planned for 1997 deployment. There are no comparable systems announced or anticipated in the same time frame. The option to procure HPSS is expected to be a cost benefit to both federal programs, universities, and commercial customers when compared with the cost of developing and integrating a system in house.

The preliminary impact to IBM, the industry, and the economy regarding jobs created/saved, revenues generated and other indirect benefits should be evident after the first twenty-four months of commercial deployment. Measurable impact should be evident after five years of commercial availability.

References

1. D. Metcalfe, and D. Thompson, "Storage Management at Cray Research, Inc.", in Cray User Group 1996 Spring Proceedings, 1996.

2. D. Teaff, R. W. Watson, and R. A. Coyne, "The architecture of the High Performance Storage System (HPSS)", in Proceedings of the Goddard Conference on Mass Storage and Technologies, March 1995.

3. R. W. Watson, and R. A. Coyne, "The Parallel I/O Architecture of the High-Performance Storage System (HPSS)", in Proceedings of the 14th IEEE Mass Storage Sympoisium, 1995.

4. Open Software Foundation, Distributed Computing Environment Version 1.0 Document Set. Open Software Foundation, 1992.

5. Internet Standard Request for Comment 959, "File Transfer Protocol (FTP)", October, 1985.

6. Technical Committee on Operating Systems of the IEEE Computer Society. IEEE Standard Portable Operating System Interface for Computer Environments, IEEE Std 1003.1-1990, January,1990.

7. R. Sandberg, et al., "Design and Implementation of the SUN Network File System", in Proceedings of USENIX Summer Conference, June, 1989.

DISTRIBUTION:

1 MS	0986	S. M. Kelly, 2665
1	0806	W. H. Rahe, 4621
1	9011	H. D. Jones, 8910
1	0661	W. D. Swartz, 4818
1	0807	R. M. Cahoon,4918
1	0807	M. L. Barnaby, 4918
5	0807	R. A. Haynes,4918
1	0188	C. E. Meyers, 4523
1	9018	Central Technical Files, 8523-2
5	0899	Technical Library, 4414
2	0619	Review & Approval Desk, 12630 For DOE/OSTI