

PROBE - A DISTRIBUTED STORAGE TESTBED

R. D. Burris, D. L. Million, S. R. White
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830

M. K. Gleicher
Gleicher Enterprises, LLC
San Diego, CA 92122

H. H. Holmes
Mass Storage Group
Lawrence Berkeley National Laboratory
Berkeley, CA 94520

I INTRODUCTION

As computers become more capable, researchers of all types are finding it necessary to store massive quantities of data generated by simulations or experiments and to retrieve them at high rate for analysis or visualization. As a consequence, strong needs have arisen for storage systems tuned for particular needs; significant improvements in storage speed and access control; optimized wide-area-network bulk transfers; database management systems capable of use with hierarchical storage systems; utilization of new media and new types of storage devices; and development, testing, and use of user-written storage applications. The Oak Ridge National Laboratory (ORNL) and the National Energy Research Scientific Computing Center (NERSC) have formed a wide-area distributed testbed -- titled "Probe" -- to support challenging storage-related studies.

"This submitted manuscript has been authored by a contractor of the U.S. government under Contract No. DE-AC05-98OR22464. Accordingly, the U. S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this Contribution, or allow others to do so, for U. S. Government purposes."

This paper explains the rationale for the creation of Probe, describes the anticipated studies that influenced the design, and shows the hardware and software configuration at the two sites. In addition, it describes completed, active, and anticipated projects, discusses the purpose of the project, results obtained and the consequences of the work and how it has been or will be applied to production operation. Briefly, projects investigate high-bandwidth transfers over wide-area networks, optimizing High Performance Storage System (HPSS) performance on various platforms and modeling (simulation) of storage processes and applications.

Probe is funded by the Department of Energy's Office of Mathematical, Information and Computational Sciences (MICS). It is available to researchers with appropriate interests regardless of affiliation, although preference will be given to interests paralleling those of MICS.

II CREATION

A. Rationale

Several areas of science, such as global climate, high-energy and nuclear physics, and bioinformatics, are generating massive quantities of data now and project huge increases in the near future, with some experiments entering the petabyte/year range. Data are being acquired from the real world and are being generated by simulation studies performed on massively parallel supercomputers.

For many applications HPSS is the software of choice for storing and managing such massive quantities of data. However, HPSS was designed to handle large numbers of huge files, while many users wish to store millions of small files. Also, HPSS implements a hierarchical storage system. Few database management systems recognize hierarchical storage. Other problems include transporting data to HPSS, finding and retrieving data appropriate to the problem at hand, and transporting data across national networks. Probe provides a facility in which such questions can be addressed.

B. Target Studies

The initial Probe studies targeted high-bandwidth visualization, wide-area distributed storage, and several activities associated with HPSS performance and utilization. These needs produced requirements for state-of-the-art communication gear, high-performance nodes, the newest disk and tape technology, and secure distributed computing software.

C. Configuration

Probe was implemented as a two-node distributed system, with one node at ORNL and the other at NERSC. The two sites are members of the ESnet network. Each has a Principal Investigator and is managed according to the policies of that site, but projects can use both sites as appropriate.

Both sites bought new IBM RS/6000 computers to act as HPSS core-server hosts and implemented separate HPSS instantiations. Both sites also implemented separate DCE cells to isolate their Probe cells from production activities. ORNL augmented the visualization equipment of their Origin 2000 supercomputer and procured each of the HPSS-supported "mover" platforms (the servers that actually control the storage equipment).

Both sites emphasize high-capacity advanced communication equipment, since storage bandwidth is so critically dependent upon the availability of big "pipes". ORNL and NERSC both bought Gigabit Ethernet and fibrechannel equipment. ORNL bought a Gigabyte System Network (GSN) switch. (GSN is also known as HiPPI 6400 because it provides 6400 megabits/second transport.) NERSC already had serial HiPPI equipment.

1. Hardware

ORNL's configuration is shown in Figure 1. Summarizing the contents, the Probe cell contains two large server nodes, redundant DCE servers, mover nodes from each supported vendor, the Origin 2000 visualization system, and various older RS/6000 computers.

NERSC's configuration is shown in Figure 2. NERSC also has redundant DCE servers, a large host for HPSS core servers, and mover nodes from IBM and Sun.

Both sites depend upon Gigabit Ethernet for the primary communication path between systems.

Figure 1.
ORNL Probe Cell, ~4/2000

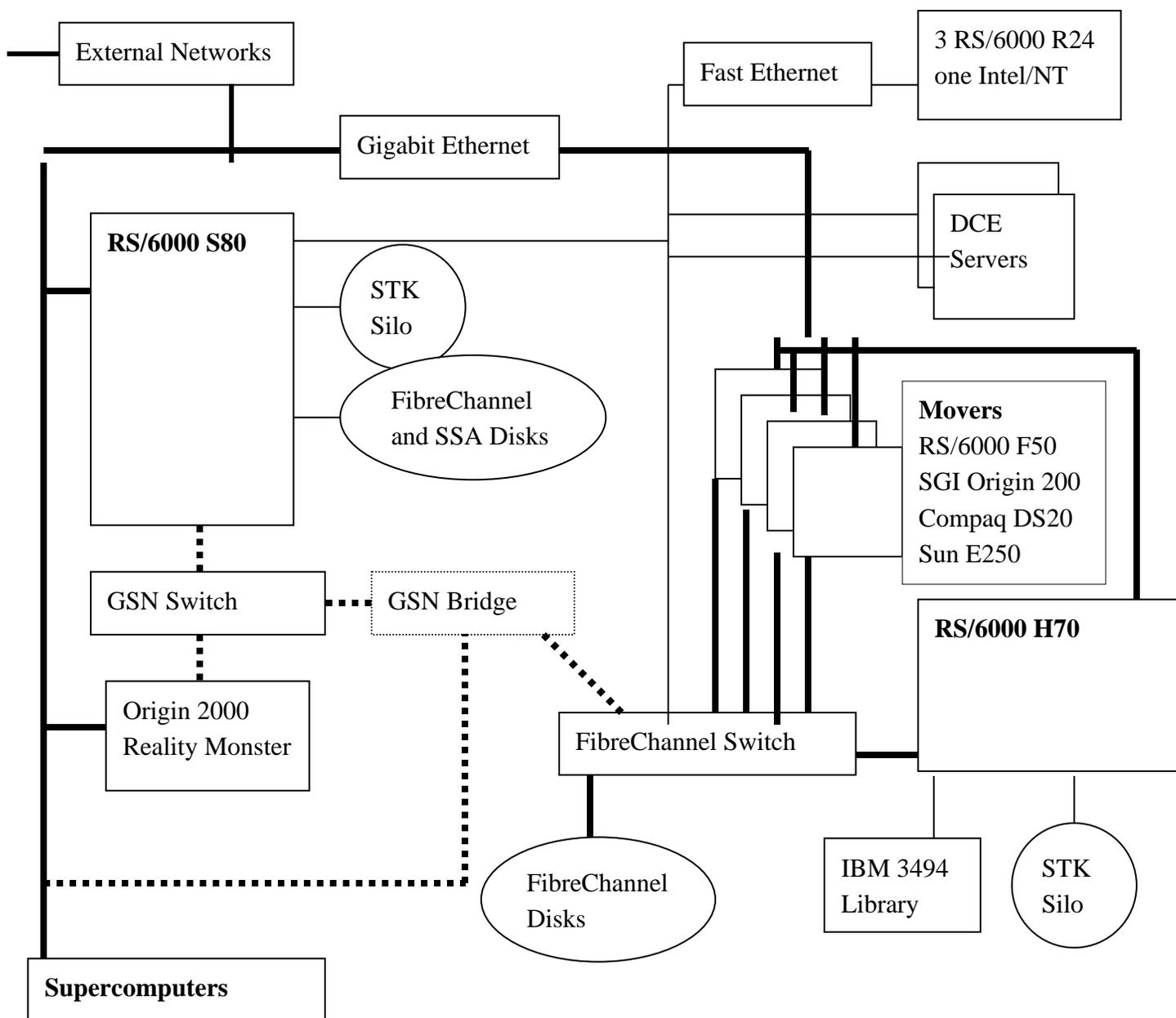
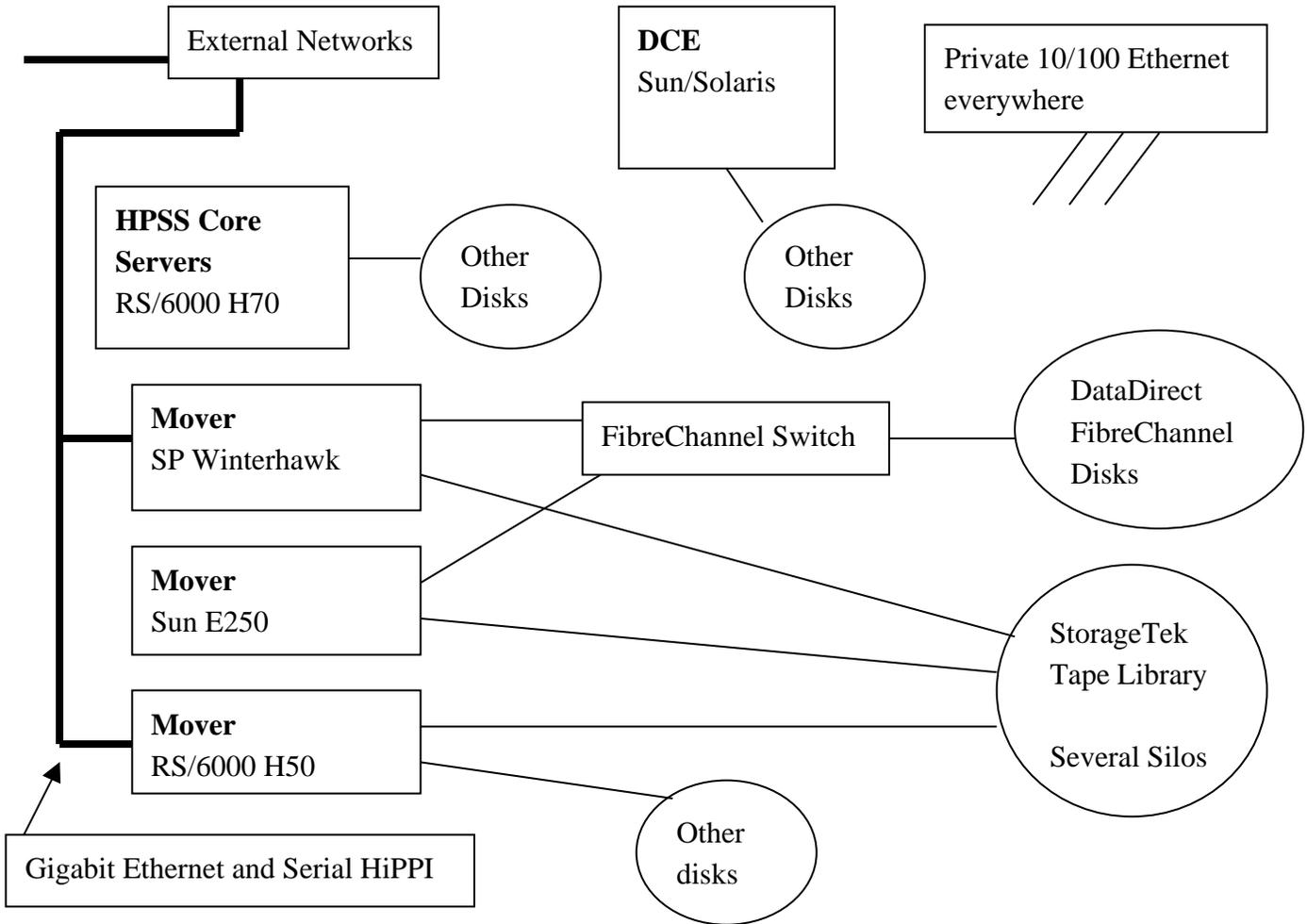


Figure 2
NERSC Configuration, ~4/2000



2. Software

Both sites are running HPSS version 4.1.1, DCE 2.2, Encina TXseries 4.2, and Hierarchical Storage Interface (HSI) software. In addition, ORNL has DB2/UDB and Oracle 8 licenses for database-related activities.

D. Rules

Projects that need to run at one installation are administered by that site. If both sites are involved, administration is shared. There are rules defining the priority to be given to a project as a function of the strength of value to the MICS office. Small projects can be performed without cost to the user; larger projects require additional funds.

III COMPLETED PROJECTS

The primary equipment at ORNL was on site in the fourth quarter of calendar year 1999. Installation and configuration of the S80 and H70 nodes was completed in mid-December. The installations at NERSC will be completed shortly.

As of late February the following projects have been completed.

A. RS/6000 S80 Performance Testing

In this project the CPU performance of ORNL's IBM RS/6000 Model S80 node was examined in response to an urgent request by the Lawrence Livermore National Laboratory. The S80 is configured with one CPU blade with six processors and two gigabytes of memory. (The machine can have up to 4 blades, 24 processors.) It has one I/O rack with fourteen PCI slots on four PCI busses. (Maximum configuration is four I/O racks with 56 PCI slots.)

The testing showed the S80 to be an outstanding integer-arithmetic engine. The full results of the test are too detailed for this paper, but they are available from the Probe web page at <http://www.csm.ornl.gov/PROBE/>. Several queries into the results of this testing have occurred.

B. Gigabit Ethernet Throughput Testing

Normal Gigabit Ethernet packets are 1500 bytes in size. Experience has shown that server nodes attempting to fill a Gigabit Ethernet pipe demonstrate very high CPU usage because of

the many interrupts required by the relatively small packet size. An expanded packet, called a jumbo frame, has been defined and is supported by some vendors

Jumbo frames were implemented between ORNL's S80 and H70 nodes. In the most extreme test, nearly 11 terabytes were transferred between the machines in a little less than 33 hours, for a sustained transfer rate of 93 million bytes/second. CPU usage on the S80 was roughly half of one processor and on the H70 was roughly 90% of a processor.

As a result of this testing, projects to investigate jumbo frames for production communications between ORNL's supercomputers have been initiated.

C. FTP Bandwidth Improvement Between ORNL and NERSC

In production transfers between the sites, users were seeing rates of 250 kilobytes/second. For ordinary transfers this rate would be sufficient, but two of ORNL's production projects have very large transfer requirements. One, global climate, occasionally has a need to ship roughly a terabyte. The other, human genome, ships five to ten gigabytes/week. The normal rates were causing long delays and problems in implementing automated transfers.

Testing was performed in Probe. The usual culprits were identified - buffers and blocks specified with default sizes to satisfy the majority of typical users. Changing settings and modifying ftp in the test environment achieved up to nine megabytes/second between the sites, the limit being FDDI links in LANs at both ends. (The sites are both on ESnet with OC3 bandwidth.) Average transfer rates were raised to over four megabytes/second; the limiting factor was congestion. This activity spawned another project, discussed below, to transfer the results to the production environment.

D. Several Smaller Projects

There have been some smaller projects, none that is earthshaking but important because they could not have been done as well or quickly had Probe not existed. In one case, a National Science Foundation (NSF) researcher at the University of Vermont wanted several gigabytes of mesh data enabling him to perform testing of a retrieval algorithm. As a result of this project the scientist got his data and Probe got relationships with the researcher, NSF, and the provider of the data, as well as a 5-gigabyte dataset that has proven useful in other activities.

In another small project, an ORNL scientist heard we had a machine in which he had some interest. He asked if he could run a benchmark test. The test itself took seconds and would

be otherwise unremarkable, except that when he heard we had other platforms as well he asked to perform the same test on those machines.

Most recently, ORNL purchased fibrechannel disks, a fibrechannel switch and interfaces for computers manufactured by four vendors. The purpose of the procurement was to establish machines for testing and tuning HPSS mover performance on the four platforms. However, we found that the installation and configuration of the switch and the computers was far from commonplace and that the records of the activity are of value to other sites. Accordingly, we have identified the work as a project and posted logs, configurations, settings, and other information.

IV ACTIVE PROJECTS

A. Improvement of Production Bandwidth Between ORNL and NERSC

As mentioned earlier, we are transferring the lessons learned in our testing to the production environment. The steps in this project include replacing the FDDI choke at both ends of the link and selecting machines in the production HPSS cells at both sites upon which a modified ftp daemon would run. Success will greatly expedite the production sharing of the global climate and human genome data.

B. Improvements to Hierarchical Storage Interface (HSI)

FTP is not the only end-user access mechanism in use at the two Probe sites. Several HPSS sites use the HSI application as the preferred end user access mechanism. Modifications to HSI are being made to investigate different ways of transferring files between two HPSS systems. The fastest would provide a true third-party transfer directly from a mover in ORNL's HPSS to a mover in NERSC's HPSS. Alternatively, data could be passed through an intermediate server at one end. The results will be valuable to any pair of sites wishing to transfer data conveniently and fast.

C. Testing of HPSS 4.1.1 on AIX Version 4.3.3

The current release of HPSS was developed on AIX 4.3.2. Release 4.3.3 is now current and includes significant changes. In this activity, HPSS has been recompiled under 4.3.3 on the S80 and has been running without problems for several weeks. Additional testing, including operation with fibrechannel disks and new tape drives (StorageTek 9840's) is pending.

ORNL plans to keep a stable generally-available version of HPSS on the S80. Success in this project will give other HPSS sites confidence in upgrading their HPSS nodes to AIX 4.3.3.

D. Testing and Tuning of HPSS Movers on SGI, Sun, Compaq, and IBM Platforms

This is the project for which the fibrechannel equipment and the Sun, Compaq, and SGI nodes were purchased at ORNL; an RS/6000 Model F50 had been purchased earlier. The equipment from each vendor was configured to have roughly the same list price (about \$20,000). Each node has 512 megabytes of memory, two disk drives, and Gigabit Ethernet and fibrechannel interfaces.

The fibrechannel equipment includes twenty 9-gigabyte drives and ten 18-gigabyte drives. Each drive is 10,000 rpm. The fibrechannel switch has sixteen ports, two that are connected to the disks and two each to the Sun, the Compaq, and the RS/6000. (The SGI has only one interface.) The drive sizes and speeds were chosen to optimize transfer rate, not capacity.

The testing is about to begin. Studies will include different RAID set sizes, different RAID levels, and various file sizes. The tests will be performed one platform at a time, with that platform having sole access to all the fibrechannel disks. We are particularly curious about RAID-3 performance for multi-gigabyte files.

Complementary testing at NERSC will be performed, using their fibrechannel disks from a different vendor, different network, and node hardware. The results of these tests will assist any sites acquiring fibrechannel equipment.

E. Modeling

A particularly important and exciting project seeks to develop a model of HPSS and storage applications. The model will be used to predict the performance of configurations that are too expensive to prototype or not yet available, to search for bottlenecks in real systems, etcetera.

To date, two commercial network simulation packages are being evaluated; purchase will be initiated within a month. At first a high-level model of the Probe system will be developed, then additional detail about HPSS servers will be incorporated over time. The model will be verified against actual Probe results, after which modifications to Probe and the model will be verified against one another.

V FUTURE

A. Planned Projects

- *ESnet III testing.* ORNL and NERSC are both members of ESnet. With installations in both sites and with the ability to saturate fast networks, Probe has been designated an "early tester" of ESnet III. When the new equipment becomes available Probe will be used to study/optimize transfers between sites and to prototype the production use of the link.
- *Supercomputer-to-supercomputer via jumbo frames.* ORNL has two large supercomputers, an IBM SP and a Compaq AlphaCluster SC. Both support Gigabit Ethernet jumbo frames through Alteon switches. Two Alteon switches have been purchased and installed. The two supercomputers will be connected, first directly and later through the switches, to investigate performance. A subsequent step will gang multiple links.
- *GSN links to visualization engine.* ORNL also has the Origin 2000 supercomputer mentioned earlier. It is capable of supporting a GSN interface, and the other supercomputers will have such interfaces available in the second quarter of 2000. ORNL has a GSN switch with six ports and a network interface card for the SGI platform. Interfaces for the other platforms will be acquired when they become available. Transfers between the supercomputers will be studied.
- *GSN bridging.* Genroco, Inc., has developed a GSN bridge with blades to aggregate eight Gigabit Ethernet circuits or eight fibrechannel links to one GSN uplink. ORNL will purchase these devices shortly and study transfers from disk to the supercomputers and to investigate Gigabit Ethernet issues.
- *HPSS performance as a function of distribution.* HPSS servers are ordinarily run on several nodes. The ORNL S80, with its six RS64-III processors, has enough CPU power and I/O bandwidth to consider running all of HPSS. It then becomes interesting to compare performance with HPSS entirely on one node, then with movers on separate nodes, then with other high-impact servers shifted to other nodes.
- *Scheduled Transfer.* A recent exciting protocol development is "scheduled transfer". This occurs when buffers are pinned in the receiving host prior to the transfer of data. This operating system bypass permits very high transfer rates unencumbered by the need to acquire buffers repeatedly and minimizing the need for memory-memory transfers.

Drivers to implement scheduled transfer will be available in the next calendar quarter and will be acquired for study at ORNL.

- *Tape drive performance comparisons.* ORNL has StorageTek 9840 and IBM 3590E tape drives. Comparisons between the two are being considered.
- *HPSS metadata performance.* NERSC will investigate HPSS performance for metadata operations and for small and large file transfers.
- *Encina file layouts.* NERSC will evaluate file layouts in the Encina Structured File System, the database system containing the HPSS metadata, to optimize metadata performance. Metadata performance is particularly important for installations with millions of files.
- *Third-party HiPPI transfers.* NERSC will perform testing of third-party transfers over serial HiPPI and fibrechannel. Success will improve the effective throughput of file transfers.

Beyond these studies, there are very interesting new technologies approaching commercialization that will be pursued as time and resources permit.

B. Possible Liaisons

A Memorandum of Understanding with the National Science Foundation is being developed. Preliminary discussions have also been held with two additional HPSS installations regarding possible involvement.

VI SUMMARY

ORNL and NERSC have established an extensive and powerful testing facility in which extremely challenging storage problems can be studied and prototyped. A variety of interesting and important projects are underway and planned, including characterization and optimization of storage and network equipment and protocols, modeling and simulation, and HPSS optimization.

The Probe facility is available to interested researchers irrespective of their affiliation, with priority going to applications most closely aligned with the interests of the Department of Energy's MICS Office.