ORNL/TM-2002/62

PROBE PROJECT STATUS AND ACCOMPLISHMENTS – Year Two

February 1, 2002

Prepared by
Randall D. Burris

DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via the U.S. Department of Energy (DOE) Information Bridge:

Web site: http://www.osti.gov/bridge

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: 703-605-6000 (1-800-553-6847)
TDD: 703-487-4639
Fax: 703-605-6900
E-mail: info@ntis.fedworld.gov
Web site: http://www.ntis.gov/support/ordernowabout.htm

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and International Nuclear Information System (INIS) representatives from the following source:

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Telephone: 865-576-8401
Fax: 865-576-5728
E-mail: reports@adonis.osti.gov
Web site: http://www.osti.gov/contact.html

ORNL/TM-2002/62

PROBE PROJECT STATUS AND ACCOMPLISHMENTS – Year Two

R. D. Burris
S. Cholia
T. H. Dunigan
F. M. Fowler
M. K. Gleicher
H. H. Holmes
N. E. Johnston
N. L. Meyer
D. L. Million
G. Ostrouchov
N. F. Samatova

February 1, 2002

Prepared by
OAK RIDGE NATIONAL LABORATORY
PO Box 2008
Oak Ridge, Tennessee, 37831-6285
Managed by
UT-Battelle, LLC
for the
U.S. DEPARTMENT OF ENERGY
Under contract DE-AC05-00OR22725

CONTENTS

Page

INTRODUCTION

The Probe project has established a facility for storage- and network-related research, development and testing. With sites at the Oak Ridge National Laboratory (ORNL) and the National Energy Research Scientific Computing Center (NERSC), Probe is investigating local-area or wide-area distributed storage issues ranging from data mining to optimizing retrieval operations from tape devices.

Probe has completed its second full year of operation. In this document we will describe the status of the project as of December 31, 2001. This year we will structure this document by category of work, rather than by project status. We will present sections describing Scientific Discovery through Advanced Computation (SciDAC) projects, network research and research on data mining and distributed cluster analysis. Another section will describe data-transfer application development and testing and other types of hardware- and software-related testing and development activities. We will then describe the work undertaken for presentation at the SC2001 conference. The final section will summarize this year's publications.

Individual projects described in this document have used some Probe resource – equipment, software, staff or funding. By describing these projects we do not imply that the work should be entirely credited to Probe, although we do assert that Probe's existence and assistance provided benefit to the work.

The Probe project is funded by the Mathematical, Information, and Computer Sciences (MICS) department of the Advanced Scientific Computing Research office, Office of Science, Department of Energy.

1. CURRENT CONFIGURATION

Appendices B and C detail the configurations of the Probe installations at the Oak Ridge National Laboratory and the National Energy Research Scientific Computing Center.

2. SCIENTIFIC DISCOVERY THROUGH ADVANCED COMPUTATION

The use of Probe facilities at ORNL was specified in several funded SciDAC proposals. Probe activities were a funded portion of the Scientific Data Management Integrated Software Infrastructure Center; Probe funding was not sought in the other projects.

2.1 SCIENTIFIC DATA MANAGEMENT INTEGRATED SOFTWARE INFRASTRUCTURE CENTER (SDM-ISIC)

Two roles for Probe resources and staff were included in this proposal. First, Probe is a "place to be" – a testbed in which other elements of the ISIC could be implemented and tested, followed by wide-area studies involving both Probe sites. The second role involves research and development into more efficient tertiary I/O. The following subsections describe current initiatives of the SDM-ISIC.

2.1.1 Probe as "A Place To Be"

Probe provides a prototyping environment for the use of other projects within the ISIC.

2.1.1.1 Hierarchical Resource Manager

ORNL/Probe has provided two nodes, a Sun machine and an IBM machine, in support of two Grid projects (see 2.2 and 2.3). In association with one of those projects, NERSC researchers have installed their Hierarchical Resource Manager (HRM), a SciDAC middleware project which is useful to the SDM-ISIC. To support the HRM work, 120 gigabytes of fibrechannel disk capacity have been attached to the Sun.

2.1.1.2 Distributed Data Analysis

With increasing frequency, researchers need access to multiple sets of data or to portions of the same set of data which exist at different sites – distributed data access and analysis. Typical data mining applications require that the entire dataset exist at one site, on one machine. This requirement cannot be fulfilled with many types of data now being analyzed, such as high-energy physics data, human genome data, climate data, etc. Consequently, research into mechanisms by which data can be analyzed without bringing all data to a single node has become important.

Two ORNL/Probe RS/6000 nodes are being used to provide support to that type of research. In support of the work, AIX licenses for Fortran (required for some of the analysis programs) and additional disk space are being provided.

2.1.1.3 Acquisition of Linux Nodes

ORNL/Probe has acquired four dual-processor Pentium III nodes for use by SDM-ISIC researchers. Initially they will be used by the projects described in 2.1.1.1 and 2.1.1.2.

2.1.1.4 Agent support

Agent technology is a crucial underpinning of the SDM-ISIC. Probe resources include a variety of platforms and good network connectivity, thus providing a solid infrastructure for agent development and testing.

2.1.2 More Efficient Tertiary I/O

A typical application retrieves an entire file, then selects from the file those data of immediate interest. Unless the entire file is of value, ignored data represent wasted resources: memory and processing power on the client, network bandwidth, memory and processing power at the data source. That waste would be reduced or eliminated if a user could specify interesting data, transport those criteria to the source of the data and implement that selection at the source.

ORNL staff are investigating ways of reducing that waste. We are looking at MPI-IO to understand how it communicates with HPSS with the intent of using its hints mechanism more effectively. We are studying HPSS code, especially “mover” code, to see how we might implement the hints at the source. We are also investigating the possibility of integrating HPSS client API calls into the NetCDF and HDF libraries to allow direct access to HPSS-resident datasets at multiple sites by application programs and visual analysis tools such as GraDS.

Ideally (but optimistically) no uninteresting data would be read from tertiary storage, transmitted or discarded. If we could achieve this ideal, retrievals would be much faster and more efficient, would have absolute minimum network impact and unburden all resources to the maximum possible degree.

2.2 EARTH SYSTEMS GRID II

The use of Probe resources was specified in the ORNL portion of this proposal. To support it, ORNL/Probe has supplied an RS/6000 Model 44P-170 running the AIX operating system and a Sun E250 running Solaris and augmented with 120 gigabytes of fibrechannel disk capacity. The Sun is being used by NERSC staff working with the Hierarchical Resource Manager (see 2.1.1.1). Argonne staff members have requested accounts so they can debug Globus software on AIX.

2.3 DOE SCIENCE GRID

The same machines as noted in 2.2, supported by the same staff at ORNL, are used in this project.

2.4 TERASCALE SUPERNOVA INITIATIVE

This SciDAC Application requires significant storage resources, processing power and network bandwidth to support visualization of the massive datasets produced in their simulations. ORNL/Probe resources will be used as researchers determine how to select, render and transport data to visualization equipment – probably across wide area links and to multiple destinations.

2.5 COMMUNITY CLIMATE SYSTEM MODEL

This SciDAC Application includes a need to transport massive quantities of climate simulation data across the wide area network. The Probe-funded enhancements of HSI, described later in this document, have made this transfer faster and easier. Additional Climate-related work utilizing higher levels of the SDM-ISIC will be implemented first in Probe, as described in the next section.

3. DATA MINING AND DISTRIBUTED CLUSTER ANALYSIS

3.1 MINING DISTRIBUTED SCIENTIFIC DATA FROM THE DESKTOP

As part of our work in DOE's Probe Project, we have developed methods requiring little communication (RACHET) that will enable scientists to perform cluster analysis of distributed data on a computational grid or on the Internet (Samatova et al., 2002). We also developed an algorithm for distributed dimension reduction by principal components (Qu et al., 2002) as an alternative method for reducing communication in RACHET. This will be incorporated into the RACHET framework, but it is of interest in itself as a method for visualization of massive distributed data sets.

Cluster analysis and dimension reduction are fundamental to discovery and visualization of structure in high-dimensional data. These computationally demanding methods are used across many data-intensive applications ranging from astrophysics to climate simulations, high energy physics experiments, and biological databases. To analyze these simulated or collected data, researchers previously required transferring large amounts of data to a central high performance computer. For massive distributed data sets, this approach is either impossible or impractical. The central idea behind distributed methods of RACHET is that a software code – not the data – is moved to a remote host that is close to the data. The code performs local analyses on local data without any communication, unlike parallel methods. The

code transfers only minimum summary information to a merger site (e.g., desktop), where these summaries are combined into a global analysis. These distributed methods are being developed for a variety of well-known data analysis techniques.

The research performed under this project will be deployed as the core of the "Multi-agent based high-dimensional cluster analysis" task in the SciDAC Scientific Data Management ISIC (PI: Arie Shoshani, LBL). New algorithms that will enable Terascale analysis of distributed and dynamically changing scientific datasets are currently being developed under the Probe project and implemented, packaged and made robust under the SDM ISIC. The application of these algorithms to the datasets generated by the SciDAC Terascale Supernova Initiative (PI: Tony Mezzacappa, ORNL) and the climate data generated by the SciDAC Community Climate System Model (PI: John Drake) are under way. This research will be leveraged under the NCSA TeraGrid initiative in FY2002 with application to computational biology research performed in collaboration with Argonne National Laboratory (PI: Natalia Maltsev).

Our current and future work includes incorporating the distributed principal components algorithm (Qu et al., 2002) into the RACHET data clustering framework. We expect this to improve clustering performance and provide a control parameter that can vary the amount of approximation to utilize available network bandwidth. We will also use the concepts developed for distributed data to design updating methodology for clustering that can efficiently deal with dynamically growing data sets.

We are working on developing algorithms for efficient merging of spanning trees. To that end, we will investigate single-linkage clustering algorithms as an alternative to the current centroid-based clustering approaches comprising RACHET's framework. Single-linkage algorithms are based on building a minimum spanning tree of the data; this work will enable single linkage clustering of distributed data with low data transfers. Because of the exceptional ability of these algorithms to deal with massive data sets, this is an important class to investigate in Probe.

3.2 CRACKING COMPUTATIONAL COMPLEXITY FOR GENOME-SCALE IN-SILICO BIOLOGY

The abundance of genomic data currently available has led to the creation of computer models of living cells. These models are essential to many applications including low-cost drug discovery, metabolic engineering, and bioremediation. However, even the simplest living cell is so complex that current supercomputers cannot simulate its behavior perfectly. The size and complexity of this problem requires the development of scalable algorithms that can take advantage of today's advances in mathematics and high performance computing. Mathematicians and computer scientists at ORNL, working in collaboration with Genetic Circuit Research Group of the University of California at San Diego (PI: Professor Bernhard Palsson), have advanced an algorithm for generating the set of extreme metabolic pathways of an organism to a scale previously not available by reducing computational time from several days to a few hours and reducing computer memory requirements by over 90%. These extreme pathways are then used to analyze, interpret, and perhaps predict metabolic functioning and control of a living cell.

The approach trades algorithm complexity for computer time and storage requirements. The result is a complex but smart algorithm that is much faster and uses less storage. It transforms a large problem into a set of small subproblems with cumulative computational cost much less than the aggregate problem. The ability to perform these subproblems almost concurrently coupled with resolution power of today's massively parallel computing platforms leaves the doors open for further improvements.

3.3 Model of a Complex Phase Transition in 2 D

A numerical modeling of a 2D phase transition has been performed and results were found to be in excellent agreement with experimental data for a class of solid surfaces. A complex phase transition in Sn/Ge(111) and similar systems can be decomposed into two intertwined phase transitions: a structural symmetry lowering ($\sqrt{3}x\sqrt{3} \leftrightarrow 3x3$) transition and a disorder-order transition in the defect distribution.

Two phenomenological models have been developed that describe these transitions and their interrelation. These models allowed us to understand the formation of domains and domain walls at low temperatures, defect induced density waves above the structural transition temperature, and ordering of the defects caused by lattice-mediated defect-defect interactions. The models predict a destruction of the pure structural transition when impurities are introduced into the system, a shift in the structural crossover temperature with impurity density, and a dependence of the 3x3 lattice structure on the specific defect alignment. The computationally intensive calculations were based on self-consistent iterative algorithms for a large two-dimensional atomic lattice and a wide range of parameters and thus utilized ORNL high performance computing resources.

4. NETWORK RESEARCH

Network research has been an important part of the work performed in association with Probe this year. In this section we will briefly describe various projects undertaken this year by Tom Dunigan and Florence Fowler of ORNL. A very extensive and informative set of Web pages developed by Tom Dunigan describe the various elements of this work http://www.csm.ornl.gov/~dunigan/. We will refer to various individual pages throughout this section.

The projects described below, together with the HSI work described in the next section, contributed to the improvement of the effective bandwidth between ORNL and NERSC by a factor of 50. The Web100 tuning work and HSI also contributed to a demonstration at SC2001.

4.1 WEB100

Several activities associated with the Web100 project received ORNL/Probe support. For more information go to http://www.csm.ornl.gov/~dunigan/netperf/web100.html.

4.1.1 Webd

We developed a simple Web100 daemon that has a configuration file of network addresses to monitor and report a selected set of Web100 variables when a "watched" stream closes. The data are recorded in a flat ASCII file suitable for statistical analysis or auto-tuning. The statistics will feed a database to be developed in the Net100 project.

4.1.2 Work-Around Daemon

We developed a prototype Work-Around Daemon (WAD) that can auto-tune the buffer sizes for designated network flows. A simple configuration file defines what remote host/port the WAD can tune and what size the send/receive buffer size should be for that flow. WAD checks for new TCP connections every second via the Web100 API and compares new connections with the configuration file to see if the flow should be tuned. Tests have been run from NERSC, ANL, SDSC, UCAR, UT, SLAC, ISDN, SLIP/PPP and home cable systems. Wide-area networks included ESnet (OC12/OC3), UT (BR/OC3) and Internet 2. Local-area networks included 100T and Gigabit Ethernet (including jumbo frames).

4.2 A TCP-OVER-UDP TEST HARNESS

The ORNL/Probe project provided support to the development of "almost TCP over UDP (atou)," an instrumented and tunable version of TCP that runs over UDP. The UDP TCP-like transport serves as a test harness for experimenting with TCP-like controls at the application level. The implementation provides optional event logs and packet traces and can provide feedback to the application to tune the transport protocol, much in the spirit of Web100 but without the attendant kernel modifications.

The experimental UDP protocol includes segment numbers, time stamps, selective ACKs, optional delayed ACKs, sliding window, timeout-retransmissions with rate-based restart, bigger initial window, bigger MSS, burst avoidance, congestion avoidance (but more aggressive, experimenting with initial window size and AIMD parameters).

For more information go to http://www.csm.ornl.gov/~dunigan/netperf/atou.html.

4.3 OPTIMIZING BULK TRANSFERS IN HIGH-DELAY/HIGH-BANDWIDTH NETWORKS

At ORNL we are interested in high-speed bulk data transfers between ORNL and NERSC over ESnet, a high-bandwidth (OC3 to OC12) and high latency (60 ms round-trip time) network in which TCP's congestion avoidance can greatly reduce throughput. We are interested in choosing buffer sizes to reduce loss and in developing more aggressive bulk transfer protocols, while still responding to congestion. We are looking at ways to monitor and tune TCP and also considering a congestion-controlled UDP (TCP friendly) that could do partial file writes to keep the buffers drained, and then fill holes as dropped packets are retransmitted. This project benefits from interaction with the Web100 and atou projects described above. For more information go to http://www.csm.ornl.gov/~dunigan/netperf/bulk.html.

5. DATA TRANSFER AND STORAGE DEVELOPMENT AND TESTING

5.1 Improve ORNL-NERSC Bandwidth

The average bandwidth between ORNL and NERSC was seen to be approximately 250 kilobytes/second, far below the peak of roughly 11 megabytes/second the hardware should allow. An initial project to find and remedy the cause was completed last year. Increasing the buffer sizes at both ends resulted in typical rates of roughly 4 megabytes/second with higher rates achieved until congestion limits were reached.

Subsequently, ESnet III equipment, with OC12 (655 megabits/second) bandwidth was installed at both Probe sites. For quite some time, observed bandwidth was far below expectations, with traffic from NERSC toward ORNL being particularly slow (roughly one megabyte/second). Extensive testing and characterization activity, together with cooperation from ESnet staff, eventually found routers that were dropping packets. Bulk transfers at 12 megabytes/second, roughly 50 times the initially observed 250 kilobyte/second rate, have since been seen. For more information see http://www.csm.ornl.gov/~dunigan/netperf/bulk.html, http://www.csm.ornl.gov/PROBE/nerscband.html, and http://hpcf.nersc.gov/storage/hpss/probe/bw.html.

5.2 HSI

HSI provides a friendly and powerful interface to HPSS (see http://www.csm.ornl.gov/PROBE/hsi.html and http://www.sdsc.edu/Storage/hsi). The author of HSI, Mike Gleicher, under contracts with ORNL and NERSC, has made extensive improvements to HSI.

- The HSI non-DCE HPSS client API library has been extended to provide the ability to communicate with multiple HPSS systems in a single session, and to switch freely between these sessions. HSI makes use of this capability and it could be extended to other interfaces. The ability within HSI to treat multiple HPSS systems as logical "drives" is a simple but powerful concept that will make it easier for researchers to make use of resources at several sites without requiring cross-cell authentication.
- I/O performance has been improved. The IPI-3 project at NERSC funded the initial I/O rewrite in HSI, and Probe-funded work has resulted in even greater performance improvements. The new "buffer pool" code results in fully double-buffered I/O (for both reads and writes) irrespective of the number of transfer threads, and decouples the HSI buffer size from the mover buffer size and the VV block size. Probe also funded the work to make use of multiple network interfaces if they are available, and to make use of restricted TCP ports at sites with firewalls.

Long-haul network performance has been improved. In addition to the continuing investigation of bottlenecks, Probe funded the changes in HSI and the non-DCE server to use multiple concurrent sockets for inter-HPSS copies.

The ability for HSI to communicate with different-release HPSS systems has been enhanced for HPSS releases 4.2 and 4.3 to allow runtime conversion of HPSS data structures that are transferred by the HSI non-DCE Client API. See http://www.csm.ornl.gov/PROBE/hsi.html.

The new HSI has been put into production at ORNL, NERSC, CalTech, University of Maryland, Indiana University, Maui High Performance Computing Center, LLNL, and the San Diego Supercomputer Center. Roughly 20 other sites use HSI as a primary user interface or for administrative functions.

5.3 COMPARATIVE PERFORMANCE-HPSS METADATA-MANAGEMENT ALTERNATIVES

The HPSS collaboration is replacing the current metadata engine, the flat-file Encina/SFS product, with a relational database management system. Prior to making that decision, the collaboration had to be confident that the replacement would not reduce performance. To research relative performance, Oracle, DB2, and SFS models of HPSS-relevant operations were tested in ORNL/Probe.

There were three associated sub-projects: to implement an externally-developed model of the HPSS file-create function on ORNL's "marlin" machine, to port that model to DB2 on marlin, and to port the model to Oracle on marlin. The same testing protocol was performed using each model. Results showed that Oracle and DB2 were comparable to one another and roughly eight times faster than Encina/SFS. As a result, replacing Encina/SFS with DB2 became the centerpiece of the next major release (Release 5.1) of HPSS. See http://www.csm.ornl.gov/PROBE/Pprojects.html.

5.4 GRID EXTENSIONS TO THE HPSS PFTP SERVER AND CLIENT

To meet the needs of the NERSC user community, NERSC has been using the Probe environment to improve access to HPSS by Grid-enabled applications. Existing Grid clients can access HPSS and improve their throughput by using a Grid-enabled HPSS server. Users with larger datasets, who need even faster throughput to HPSS, can install and use a new Grid-enabled PFTP client.

NERSC staff have made extensions to the HPSS PFTP server (based on work done at SDSC) to accept Grid credentials. Another enhancement to the server accepts commands from an existing Grid client to set the TCP network buffer size. Performance can be improved by setting the buffer size based on the size of the dataset and the network topology.

On the client side, NERSC has modified the PFTP client to support HPSS in Grid environments by adding Grid authentication. A PFTP client is important because data is transferred directly from a mover to a client bypassing the FTP daemon. Additionally, PFTP provides the ability to transfer data over multiple parallel streams.

NERSC tested the products with a variety of users. For example, PNNL researchers are moving data between Probe's HPSS system and computational servers in Washington. NERSC has also been working with users from the Earth Science Grid (NCAR) to move data from systems in Colorado to Probe.
The Grid-enabled PFTP server and client are merely interim measures – in the long run, HPSS will be revised to support the use of parallel transfers in a manner compatible with the Grid scheme. Both Probe sites have expressed interest in participating in such a project.

5.5 HPSS MOVERS USING GIGABIT ETHERNET NETWORK CONNECTIVITY AND FIBRECHANNEL DISKS

ORNL purchased servers from IBM, Compaq, SGI, and Sun with the goal of testing/tuning HPSS mover operation using FibreChannel disks and Gigabit Ethernet network interfaces. Some testing was performed in the first year of Probe operation.

During the second year of operation, RAID 3 tests were performed and compared to RAID 5 results. Also, mover software developed and provided by Jean-Pierre Thibonnier of Compaq was installed and tested on the Probe Compaq Alpha DS20 node. The Compaq software was easy to install and worked flawlessly.

All testing associated with this project has been completed. Results were presented to the HPSS User Forum in June 2001. See http://www.csm.ornl.gov/PROBE/commodity.html.

5.6 LINEAR TAPE OPEN (LTO) HPSS DEVELOPMENT

NERSC, in conjunction with IBM integrated the LTO system into HPSS, including developing a new LTO Physical Volume Repository and modifications to SSM and the mover. That capability was released in HPSS version 4.3.

Work is described at http://hpcf.nersc.gov/storage/hpss/probe/LTO/IBM_LTO_test_1.pdf.

5.7 REMOTE MOVERS

The "remote mover" concept describes an HPSS installation that includes a mover node at a remote site. NERSC and ORNL have tested two configurations – one in which an ORNL node is part of a NERSC HPSS installation and the converse – a NERSC node is part of an ORNL HPSS installation. We have also established a configuration in which a single node hosts movers for both installations. The benefit of the remote mover concept is that files are transferred between the sites under the control of HPSS software as a "migration" from one level of storage to a lower (and remote) level. The user does not have to wait for the transfer to complete – it takes place "behind the scenes".

The tests have been successful and have shown that the concept is valuable. At the moment the two production HPSS installations are at different HPSS releases, which may lead to operational difficulties, so we have agreed to wait until we are both at the same HPSS release level to revisit deploying remote movers in production.

5.8 TESTING AT THE REQUEST OF THE HPSS COLLABORATION

Several activities were performed to assist in the testing and support of HPSS. In one series, ORNL's StorageTek Redwood tape drives (which are unavailable in IBM/HPSS's Houston testbed) were used to test and validate HPSS version 4.2. In another, ORNL/Probe equipment was used to test HPSS version 4.3 on AIX and Solaris with StorageTek 9840 tape drives. In both cases IBM staff performed the tests. In tests of this nature, no earth-shaking conclusions are produced. In each case the testing was successful and the product has been released. See http://www.csm.ornl.gov/PROBE/P2projects.html.

5.9 FRONT-END BUFFER

Another way in which a user can cause file transfers without waiting for the transfer to complete is to implement a "spooling" capability. ORNL is developing such an ability. As designed, a user will issue an HSI command, either interactively or in batch, and that command will be communicated to and executed by a separate server. We see two benefits to this approach. One, as described earlier, will free the user from waiting for a transfer to take place. The other benefit is that retrievals, or transmissions, would not require that HPSS be available at the time the request is made. This will disassociate, to a greater degree, the maintenance schedules of the supercomputers and the HPSS system, leading to greater production reliability.

5.10 HPSS COMPATIBILITY WITH NEW INFRASTRUCTURE PRODUCTS

HPSS is tested on a specific set of infrastructure products (including DCE, DFS, Encina, Encina's SFS, and Sammi) and on two platforms, IBM/AIX and Sun/Solaris. The various products have different release schedules, so it is usually the case that soon after HPSS is released, some infrastructure product comes out with a new release. The HPSS test team has its hands full testing the functionality of new HPSS patches and releases. They cannot test/certify all combinations of HPSS and infrastructure product releases.

In an ongoing activity, ORNL/Probe instantiates the latest release of HPSS over the latest releases of infrastructure products. This involves compiling, building and running HPSS in the new environment. Success provides HPSS customers with some confidence that HPSS operates correctly with the later infrastructure. This activity has tested HPSS 4.2 over DCE 3.1and HPSS 4.3 over AIX 5.1.

5.11 HIGH-PERFORMANCE VISUALIZATION

One of the primary motivations for the creation of Probe is the investigation of high-bandwidth transfers from storage to visualization systems. ORNL undertook Gigabyte System Network (GSN) and Scheduled Transfer (ST) studies to develop and test a mechanism for such transfers. At this writing the GSN switch has been installed and connected to the Origin 2000 Reality Monster. A project to visualize the results of a simulation of a supernova explosion - using Probe servers, storage resources, and the GSN equipment - has begun. See http://www.csm.ornl.gov/PROBE/Pprojects.html for details.

5.12 MODELING CACHE PERFORMANCE IN HPSS

NERSC assembled 18 months of transfer logs from one of their production HPSS systems and analyzed them to assess workload behavior and gain some insight into which cache configurations would provide the best service to the users.

We found, as expected, that the workload is distributed over file size with a declining number of files as the files get larger, so the amount of space consumed per file size increment is roughly constant up to file sizes of 1 GB. Sixty one percent of file accesses were write accesses. There are a significant number of files written which are never read – backup files and similar files. For all sizes of files, access frequencies decline with the age of the files.

HPSS uses the cache as an I/O buffer for incoming data. At NERSC the cache behavior is dominated by the write traffic. Cache lifetimes tend to scale linearly with the size of the cache and inversely with the amount of data flow.

There is a paper on the web: http://hpcf.nersc.gov/storage/hpss/probe/caching/cache-behavior.pdf "Exploration of Cache Behavior Using HPSS Per-File Transfer Logs."

5.13 GUI INTERFACES TO STORAGE

NERSC Probe staff has evaluated a file-caching web server operated by physics and nuclear science users at NERSC. This server provides web-based interactive access to a large set of files widely used by physics and nuclear science researchers. The GUI provides a convenient content-oriented view of the data, with convenient point-and-click selection of files and downloading. Such a facility also makes possible multiple organizations of such file collections and convenient annotations regarding the files. They are participating in the operation of this server to evaluate approaches to a more general capability.

5.14 MODELING STORAGE

The acquisition, storage and use of terabytes of data requires hundreds of pieces of equipment and very complex applications. Intuition is of limited value in establishing optimal and cost-effective configurations and procedures. ORNL has established a project to develop a model of the entire storage scenario, from acquisition through analysis, first modeling HPSS. Various data sources and analyses (high-energy physics experiments, for instance) could be added as additional projects.

At this time a network modeling tool, OPNET, has been purchased and installed. Discussions with various possible sources of data within IBM and StorageTek have been held. ORNL staff will address the acquisition of performance data from various HPSS and operating-system sources. IBM/HPSS is cooperating by providing performance data and an IBM HPSS developer will be participating as time permits. A student at the University of North Dakota, Aric Broeking, and his advisor, Thomas Wiggin, have begun developing the model as Aric's Master's Thesis; other students of Dr. Wiggin are developing models of other storage entities.

5.15 SCHEDULED TRANSFER (ST)

ST is a software technology that bypasses much of the operating-system processing ordinarily performed in high-bandwidth transfers. ORNL acquired three ST licenses from Genroco for installation on the two Compaq AlphaServer SC supercomputers and the Probe Compaq DS20 server being used in HPSS mover testing.

Research has dimmed hopes of making effective use of Scheduled Transfer between heterogeneous nodes. One problem is that the current specification has been implemented only on SGI Origin equipment. A second problem is that the protocol is very light-weight, with very little error correction, so it is not appropriate for other than local-area networks.

A third problem is cost-effectiveness. Studies have shown only minimal throughput gains when compared with Gigabit Ethernet jumbo frames. Because the code is quite difficult to implement, clear and significant performance gains would be necessary to justify the effort to develop ST applications. Those performance gains do not appear likely. Accordingly we have decided not to continue work with ST.

5.16 EQUIPMENT TESTING

5.16.1 Texas Memory Systems

ORNL and NERSC participated in testing of the Texas Memory Systems RAM-SAN product at the joint request of the vendor and/HPSS. The equipment was tested for transparency (i.e., did it appear to be a normal disk to the operating system; it did), for use with HPSS's metadata processing and for performance.

At ORNL tests studied raw I/O performance of the device, verified its compatibility with HPSS, studied performance with a rotating disk mirroring the I/O to the RAM-SAN and re-ran the DBMS testing described in 5.3 above. See http://www.csm.ornl.gov/PROBE/TMS_ORNL.html for more information.

NERSC ran three different benchmarks. Initial baseline timing benchmarks were run using the UNIX utility dd. To benchmark transactional performance NERSC used the Encina database system that is used by HPSS. The final benchmarks used HPSS from the Parallel Distributed Systems Facility (PDSF) system across jumbo frame Gigabit Ethernet. Results of the three benchmarks are presented at http://hpcf.nersc.gov/storage/hpss/probe/tms/index.html.

The device performed flawlessly at both sites. In the end, it had little performance advantage over rotating disk in HPSS metadata processing, demonstrating that the bottleneck in that application is something other than disk latency. No site has purchased the device for production SFS use.

5.16.2 SCSI-FibreChannel Bridge Testing

ORNL has eight IBM 3590E SCSI tape drives and a need to connect them to a FibreChannel interface for transfer-rate and packaging reasons. To that end a SCSI-FibreChannel Bridge was acquired and used to connect two 3590E drives to a Probe HPSS node. After successful testing in Probe, all eight drives were connected to the Bridge and thence to the production HPSS installation. At that time it became possible to retire the obsolete IBM RS/6000 MicroChannel nodes to which the drives had been connected, resulting in a significant savings in maintenance costs.

5.16.3 StorageNet 6000 (SN6000)

Internal transfers within HPSS can benefit significantly from FibreChannel to FibreChannel data transfers (for instance, between FibreChannel disk and tape units). NERSC has been testing the initial version of the SN6000 unit from StorageTek as an entry point into this area. Currently five FibreChannel tape drives are attached to the SN6000 and then to two hosts. Timing tests, heavy-load tests and stability tests have been performed using different configuration options. When the SN6000 supports disk, further tests will be conducted, hopefully third party transfers.

5.16.4 RAIT/RAIL

At ORNL an SN6000 was bought to test the Redundant Array of Independent Tape (RAIT) and Redundant Array of Independent Libraries (RAIL) facility developed under the ASCI PathForward project. To date that facility has not been completely implemented and is not yet available even in beta form.

5.16.5 Storage Area Networks (SAN)/IP

The work to investigate bridging FiberChannel to Gigabit Ethernet is a preliminary to studying SANs using IP. Early products are just entering the market; some are tuned for local area networks and others for wide-area networks. They hold considerable promise for wide-area storage transfers and for less-expensive SANs. NERSC has done some preliminary tests on a loaner iSCSI box from Cisco. Transfers on a dedicated private network and some initial wide-area-network tests have been conducted.

5.16.6 Linear Tape Open (LTO) Test and Development

The PROBE testbed at NERSC had a beta test agreement for the new IBM 3584 tape library with LTO tape technology. The goal of this beta test was to assess the operation of the library and drives with AIX version 4.3.3, including performance and load tests for the LTO tape drives and the library. The testing was performed at the request of IBM and by an IBM employee subcontracted to NERSC. This project is described at http://hpcf.nersc.gov/storage/hpss/probe/LTO/IBM_LTO_test_1.pdf.

5.16.7 FibreChannel Disk Array

NERSC ran series of timing tests on a StorageTek 9176 fibre disk array. Of the various configuration parameters investigated, two were found most important – the storage array cache block size and the number of controllers and buses. Results include showing the original data and comparison graphs. The configuration of the various components and how the tests were conducted are also shown.

There is a paper on the web: http://hpcf.nersc.gov/storage/hpss/probe/timing_stk/timing_stk.html "Analysis of StorageTek 9176 Fibre Channel Disk Array."

5.16.8 Gigabyte System Network (GSN) Hardware and Drivers

Genroco has built network interface cards for several platforms (IBM, Compaq, and Sun) and operating-system software ("drivers") for each. With ORNL, they tested the IBM hardware and drivers for the RS/6000 Model S80 (finding and correcting some bugs). In calendar year 2000 a penultimate version of the interface card demonstrated a record transfer rate exceeding 150 megabytes/second between the S80 and an SGI Origin 2000. In 2001 a rate of 193 megabytes/second was obtained between an RS/6000 Model B80 and the Origin 2000. Also in 2001 ORNL tested TCP/IP between Compaq Tru64 version 5.1 and the Origin 2000.

In all tests performed involving platforms other than the Origin 2000, the node had a PCI bus. The S80 has 33 MHz slots; the B80 has 50 MHz slots. ORNL obtained a model p660-6H1, which has 66 MHz slots, expected to see an additional rise in transfer rate. However, that rise was not observed.

Subsequent discussions determined that the p660 I/O architecture was designed for very high aggregate throughput, not maximum "burst" (single-channel) performance, and in fact the maximum burst rate that could be achieved would be roughly 200 megabytes/second. There are very few interfaces (other than GSN) that support the 400+ megabytes/second that could be supported by a 64-bit 66 MHz bus, so the

design approach is understandable. As a general result of our testing experiences, however, we doubt that GSN will ever be a cost-effective communication mechanism involving heterogeneous platforms.

6. SC2001 ACTIVITIES

6.1 INTER-HPSS FILE TRANSFER DEMONSTRATION

The ORNL booth included a demonstration of the use of HSI in its HPSS-HPSS mode. Temporary accounts were established at ORNL, NERSC, SDSC and the Indiana University and a set of files stored in each location. Using the easy-to-use syntax – essentially that of logical disk notation – transfers were demonstrated between any two HPSS installations, in either direction. As an example, a transfer of file ABC from ORNL to SDSC would have been initiated by "cp O:ABC S:", where O: and S: represent ORNL and SDSC respectively.

6.2 HPSS WIDE-AREA REMOTE-MOVER DEMONSTRATION

NERSC established a demonstration of the use of HPSS Wide-Area Remote Movers. For this activity, two mover nodes were established at locations remote from NERSC – one at LBNL and one at Oak Ridge. Following the standard procedure, files were stored in the NERSC HPSS installation from the Oak Ridge HPSS installation by copying them (using HPSS-HPSS features of HSI) to a special Class of Service which caused the files to be cached on the Oak Ridge node of the NERSC HPSS. Files then migrated to a second level of disk physically sited at NERSC. Transfers were unusually fast as the processing used three parallel stripes. The advantage of this procedure is that users do not need to wait for the wide-area transfer; it is handled in the background by HPSS.

6.3 WEB100-TUNED WIDE-AREA BULK FILE TRANSFER DEMONSTRATION

Web100 functionality was used to tune wide-area bulk transfers from NERSC to ORNL. The HSI application was used in the transfers. Two versions were available; one used the standard mover-mover protocol and an experimental version eliminated some of the handshaking characterizing that protocol. The end nodes used the linux operating system. The GUI Web100 interface was used to dynamically – and in real time – tune window sizes so the observer could witness the effect on transfer rates.

6.4 PARALLEL OUT-OF-CORE ENUMERATION OF EXTREME METABOLIC PATHWAYS DEMONSTRATION

A highly scalable algorithm that enumerates all the extreme metabolic pathways on a genome-scale was demonstrated on a cluster of Linux PCs. The generation process for a selected example took on the order of fifteen minutes on four processors by our algorithm as opposed to five days by the original code obtained from the UCSD. For each iteration, the CPU time and memory requirements were displayed for both algorithms, demonstrating an improvement of computational time and memory by several orders of magnitude. With such a scalable algorithm, the generation of all the extreme pathways for the entire

organism will become possible. These pathways are then used to analyze, interpret, and predict metabolic functioning and control of a living cell.

6.5 GRID-ENABLED PFTP TRANSFER DEMONSTRATION

Section 5.4, "Grid Extensions to the HPSS PFTP Server and Client," describe work done at NERSC. At the NERSC booth, transfers from PNNL and NCAR demonstrated the capability.

7. PUBLICATIONS AND PRESENTATIONS

7.1 PUBLICATIONS & PATENTS:

Refereed Papers:

1) N. F. Samatova, T. E. Potok, M. R. Leuze (2001). "Vector Space Model for the Generalized Part Families Formation", *Robotics and CIM,* 17: 73-80 (invited paper).
2) N. F. Samatova, G. Ostrouchov, A. Geist, A. Melechko (2002). "RACHET: An Efficient Cover-Based Merging of Clustering Hierarchies from Distributed Datasets," *Special Issue on Parallel and Distributed Data Mining*, *International Journal of Distributed and Parallel Databases: An International Journal*, Volume 11, No. 2, March 2002.
3) A. V. Melechko, M. V. Simkin, N. F. Samatova, J. Braun, W. Plummer (2001). "Intertwined CDW and defect ordering phase transitions in a 2D system," *Physical Review B,* Volume 64, No. 235424.

Conference Proceedings:

1) N. F. Samatova, A. Geist, G. Ostrouchov and A. Melechko (2002). "Parallel Out-of-core Algorithm for Genome-Scale Enumeration of Metabolic Systemic Pathways," *Proceedings of the 1st Workshop on High Performance Computational Biology*, Marriott Marina, Fort Lauderdale, Florida, April, 2002 (accepted for publication).
2) N. F. Samatova, G. Ostrouchov, A. Geist, A. V. Melechko (2001). "RACHET: A New Algorithm for Mining Multi-dimensional Distributed Datasets," *Proceeding of the SIAM Third Workshop on Mining Scientific Datasets*, Chicago, IL, April 2001.
3) T. E. Potok, N. D. Ivezic, N. F. Samatova (2001). "Agent-based architecture for flexible lean cell design, analysis, and evaluation," *Proceedings of the 4th Design of Information Infrastructure Systems for Manufacturing Conference,* Melbourne, Australia.
4) T. E. Potok, M. T. Elmore, J. W. Reed, and N. F. Samatova "An Ontology-based HTML to XML Conversion using Intelligent Agents," *Hawaii International Conference of System Sciences*, 2001.
5) Harvard Holmes, "Exploration of Cache Behavior Using HPSS Per-File Transfer Logs," Lawrence Berkeley, LBNL-49330, November 2001, http://hpcf.nersc.gov/storage/hpss/probe/caching/cache-behavior.pdf.
6) S. Cholia, N. Meyer, "A Beta Test of Linear Tape-Open (LTO) Ultrium Data Storage Technology," Lawrence Berkeley National Laboratory report LBNL-49327, December 2001; http://hpcf.nersc.gov/storage/hpss/probe/LTO/IBM_LTO_test_1.pdf.

Submitted/Written Papers:

1) Y. Qu, G. Ostrouchov, N. F. Samatova, A. Geist (2002) "Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets," The Second SIAM International Conference on Data Mining, April 2002, submitted.
2) N. F. Samatova, T. E. Potok, A. V. Melechko, M. R. Leuze (2001) "On the vector space model in manufacturing cell formation," *International Journal of Production Research,* submitted.

3) N. F. Samatova, T. E. Potok, A. V. Melechko (2001) "Data Representation Taxonomy for Manufacturing Cell Formation Models," *OMEGA: the International Journal of Management Science,* submitted.
4) J. W. Reed, T. E. Potok, N. F. Samatova, and M. T. Elmore, "Dynamic Cluster Analysis for Scalable Data Mining," submitted to the *Hawaii International Conference of System Sciences*, 2001.
5) T. H. Dunigan, F. M. Fowler, "A TCP-over-UDP Test Harness," in preparation.

Patents:

1) M. T. Elmore, J. W.Reed, T. E. Potok, N. F. Samatova, J. N. Treadwell (2002). "A Process of Gathering and Summarizing Internet Information," a patent submitted to the intellectual property department.

7.2 Presentations, Posters and Demonstrations

1) N. F. Samatova, G. Ostrouchov, A. Geist, A. Melechko, "RACHET: A New Algorithm for Mining Multi-dimensional Distributed Datasets." Presented at the *SIAM Third Workshop on Mining Scientific Datasets*, Chicago, IL, April 5-7, 2001.
2) N. F. Samatova and G. Ostrouchov, "Multi-agent based High-Dimensional Cluster Analysis," SDM-ISIC Kick-off meeting (with DOE program manager and other laboratories in attendance), July 10, 2001.
3) N. F. Samatova, G. Ostrouchov, and Y. Qu, "Solution Space Characterization," presented at UCSD to the Genetic Circuits Group of the Department of Bioengineering, July 12, 2001.
4) B. Palsson, "Predictive models of biochemical pathways and microbial behavior," GTL/DOE Workshop, August 7, 2001 (minor contribution).
5) N. F. Samatova and G. Ostrouchov, "Multi-agent based High-Dimensional Cluster Analysis," presented to Steve Eckstrand, OS/DOE, August 9, 2001.
6) N. F. Samatova, "Vector Space Model for Lean Cell Formation," presented at CSMD seminar, December 1, 2000.
7) N. F. Samatova and G. Ostrouchov, "Scientific Data Mining Research under Probe," presented to Distributed Computing Group at CSMD/ORNL, September 18, 2001.
8) N. F. Samatova, G. Ostrouchov, A. Geist, B. Palsson, N. Price, J. Papin, S. Smith, "Cracking Computational Complexity for Genome Scale Modeling of Metabolic Pathways," Presented to Network and Cluster Computing Group at CSM/ORNL, October 30, 2001.
9) N. F. Samatova, G. Ostrouchov, A. Geist, "Scientific Data Mining Research at CSM," presented to Life Sciences Division, ORNL, January 11, 2002.
10) SC2001 Poster, "Rachet: Petascale Distributed Data Analysis Suite."
11) SC2001 Poster, "Cracking Computational Complexity for Genome Scale Modeling of Metabolic Pathways."
12) SC2001 Demonstration, "Cracking Computational Complexity for Genome Scale Modeling of Metabolic Pathways."
13) E. W. Plummer, A. V. Melechko, M. Simkin, N. F. Samatova, J. Braun (2001), "Medard W. Welch Award Lecture: Intertwined Charge Density Wave and Defect-Ordering Phase Transitions in a 2-D System," Presented at IUVSTA 15th International Vacuum Congress (IVC-15), AVS 48th International Symposium (AVS-48), 11th International Conference on Solid Surfaces (ICSS-11), San Francisco, October, 2001.
14) M. K. Gleicher, "Texas Memory Systems Testing," presented to the HPSS Users Forum, June 2001.
15) M. K. Gleicher, "HSI," presentation to the HPSS Users Forum, June 2001.
16) R. D. Burris, "ORNL/Probe Performance Tests," presentation to the HPSS Users Forum, June 2001.
17) R. D. Burris, D. L. Million, "Probe Plans and Status," presentation to the SciDAC SDM ISIC Kickoff, July 2001.
18) R. D. Burris, "Probe Data Storage, Transfer and Research Facility," presentation to networking workshop sponsored by Thomas Ndousse, December 2001.

19) T. H. Dunigan, "Net100 overview," Web100 conference, Boulder, CO, July 2001.
20) T. H. Dunigan, "Web100 testing at ORNL," Web100 conference, Boulder, CO, July 2001.
21) T. H. Dunigan, "Net100 project," Network BOF, SC 2001, Denver, November 2001.
22) T. H. Dunigan, "Net100," DOE network workshop, Oak Ridge, November 2001.
23) T. H. Dunigan, "Net100," DOE SciDAC PI meeting, Wash. DC, January 2002.
24) T. H. Dunigan, "Net100 measurement and tuning," Internet2 workshop, Tempe, January 2002.

8. SUMMARY

Probe has evolved from a simple testbed, early in its life, to a productive research facility with notable accomplishments in both networking and data-related science. A variety of the tests performed in Probe, and many of its development projects, have resulted in tools that have been put into production use at ORNL, NERSC and other HPSS sites around the world. It is a key component of a variety of MICS-funded activities, including both base funded and SciDAC projects. SciDAC involvement includes the SDM ISIC, the Terascale Supernova Initiative and the DOE Science Grid and Earth Systems Grid II projects.

The new focus on SciDAC activities has shifted the direction of facility matters. Prior to SciDAC we concentrated on state-of-the-art equipment; now we must concentrate on new equipment that can be reliably and quickly put into production. Existing equipment will be maintained and refreshed as necessary, but new procurements will focus on providing storage resources to support large-data activities of various SciDAC applications.

APPENDIX A
COLLABORATIONS

A.1 Genroco

ORNL has established a collaborative relationship with Genroco, maker and vendor of GSN hardware and ST software. Genroco is also prominent in SAN/IP activities, bridging various high-bandwidth network technologies to GSN, use of ST over ATM, and soon 10-gigabit Ethernet. ORNL gains maintenance benefits for GSN software and hardware in return for beta testing the products.

A.2 IBM and StorageTek

ORNL and NERSC have collaborative relationships with IBM in the hardware and software (HPSS) arena. ORNL also has a collaborative relationship with StorageTek. In each case the collaboration allows beta testing of new equipment and capabilities.

A.3 Consensys

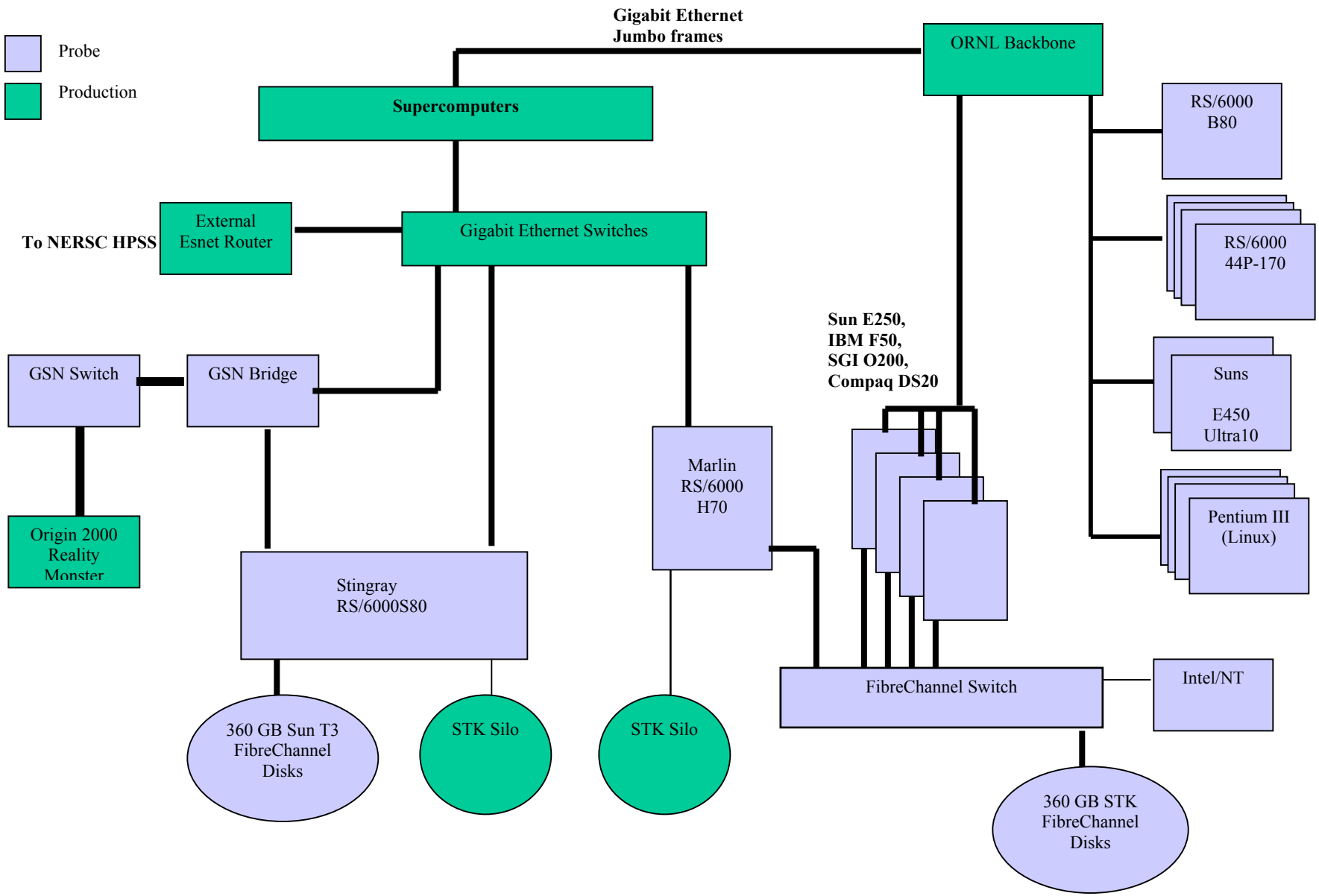
ORNL has a relationship with Consensys, makers of RAIDzone, a NFS server product, in which ORNL provides a jumbo-frame LAN and non-Linux NFS clients and Consensys provides a loaner machine and engineering staff involvement.

A.4 Brookhaven National Laboratory (BNL)

NERSC and BNL have collaborative relationships in the area of High Energy and Nuclear Physics data bases and data processing. One area of investigation is query optimization, which incorporates knowledge of file locations on the physical tapes in the HPSS storage systems. Data retrieval for multiple queries can be combined, and optimum ordering of file retrievals can speed up tape processing. This research is also looking at optimizing the original placement of data on tape, given knowledge about typical queries.

APPENDIX B
CONFIGURATION OF ORNL EQUIPMENT AS OF DECEMBER, 2001

APPENDIX B CONFIGURATION OF ORNL EQUIPMENT AS OF DECEMBER, 2001

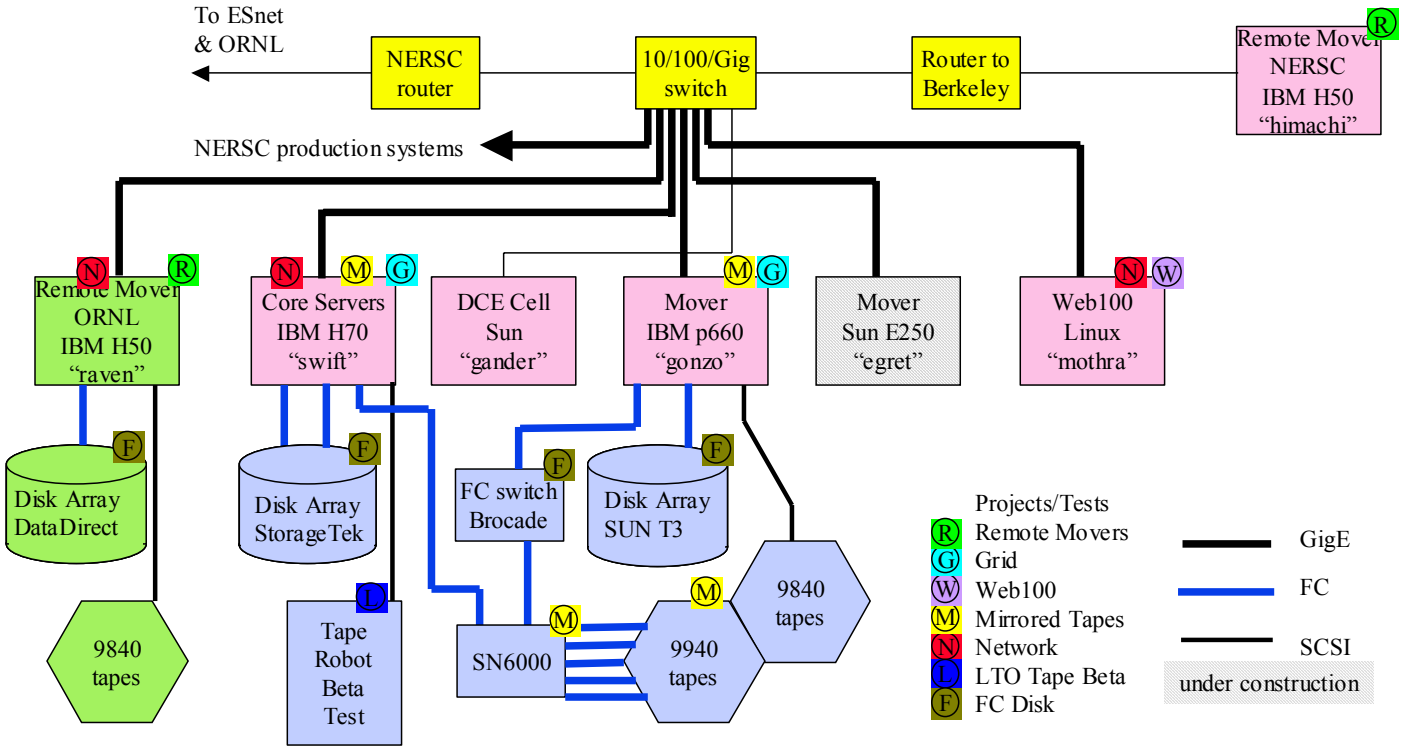


ORNL EQUIPMENT DESCRIPTIONS

Node Name	Machine	OS	Processors	Memory MB	Disk GB	Network	Software
stingray	IBM S80	AIX	6	2048	384	GigE	C, gcc, Fortran, NetCDF, HDF5 ggobi, R
marlin	IBM H70	AIX	4	2048	280	GigE	C, Fortran, DB2, Oracle
earl	IBM B80	AIX	2	1024	36	FastE	C
satchel	IBM 44P-170	AIX	1	1024	54	FastE	C
sneezy	IBM 44P-170	AIX	1	512	310	FastE	C
bucky	IBM 44P-170	AIX	1	1024	54	FastE	C
bashful	IBM 44P-170	AIX	1	512	27	FastE	C
maverick	IBM F50	AIX	2	512	133	GigE	C
happy	Sun E450	Solaris	1	512	108	FastE	C/C++ OPNET
sleepy	Sun E250	Solaris	2	512	430	GigE	C/C++, HRM
mustang	Sun Ultra10	Solaris	1	128	18	FastE	C/C++
grumpy	SGI O200	IRIX	1	512	18	GigE	C
dopey	Compaq DS20	Tru64	2	512	18	GigE	C
dilbert	Intel	Linux	2	512	240	FastE	C
wally	Intel	Linux	2	512	240	FastE	C
alice	Intel	Linux	2	512	240	FastE	C
phb	Intel	Linux	2	512	240	FastE	C
doc	Intel	NT	1				Misc.

APPENDIX C
CONFIGURATION OF NERSC EQUIPMENT AS OF JANUARY 2001

PROBE: NERSC Configuration 12/2001



NERSC EQUIPMENT DESCRIPTIONS

Node Name	Machine	OS	Processors	Memory MB	Disk GB	Network
Swift	IBM H70	AIX	4	1024	45+	GigE
Raven	IBM H50	AIX	4	1024	27+	GigE
Gonzo	IBM p660	AIX	4	1024	36+	GigE
Eagle	IBM H50	AIX	4	768	9+	GigE
Gander	Sun	Solaris	1			FastE
Egret	Sun E250	Solaris				GigE
Mothra	Intel	Linux				GigE

Note: "+" in the Disk column denotes external FibreChannel or SSA disk capacity.

INTERNAL DISTRIBUTION

1. R. A. Alexander
2. M.W. Arnold
3. A.S. Bland
4. M. Chen
5. T. Dunigan
6. R.A. Fahey
7. M. R. Fahey
8. F. Fowler
9. A. Geist
10. C. Halloy
11. N. R. Hathaway
12. R. A. McCord
13. D. L. Million
14. G. Ostrouchov
15. N. S. Rao
16. F. Samatova
17. D. A. Steinert
18. R. J. Toedte
19. V. L. White
20. S. R. White
21. J. B. White, III
22. W. R. Wing
23. B. A. Worley
24. T. Zacharia
25. Central Research Library
26. ORNL Laboratory Records (RC)
27. ORNL Laboratory Records - OSTI

EXTERNAL DISTRIBUTION

1. John Blaylock, jwb@lanl.gov
2. Jeff Bongianino, jeff.bongianino@gdesystems.com
3. Aric Broeking, broeking@scc.und.edu
4. Kasidit Chanchio, chanchiok@ornl.gov
5. Shreyas Cholia, SCholia@lbl.gov
6. Danny Cook, dpc@lanl.gov
7. Bob Coyne, coyne@us.ibm.com
8. Mike Devaney, dm_devaney@pnl.gov
9. Keith Fitzgerald, kfitz@llnl.gov
10. Jim Fox, fox@cac.washington.edu
11. Mark Gary, mgary@llnl.gov
12. Krzysztof Genser, genser@fnal.gov
13. Mike Gleicher, mkg@san.rr.com
14. Otis Graf, ofgraf@clearlake.ibm.com
15. Phil Greene, philip_greene_jr@storagetek.com
16. Bill Gropp, gropp@mcs.anl.gov
17. Annette Hamala, ahamala@us.ibm.com
18. Dan Hitchcock, daniel.hitchcock@science.doe.gov
19. Harvard Holmes, hhholmes@lbl.gov
20. Fred Johnson, fjohnson@er.doe.gov
21. Nancy Johnston, nejohnston@lbl.gov
22. William Johnston, WEJohnston@lbl.gov
23. Hilary Jones, hilary@ca.sandia.gov
24. Jae Kerr, jrkerr@us.ibm.com
25. Bill Kramer, WTKramer@lbl.gov
26. Greg Lefelar, lefelagr@us.ibm.com
27. C William McCurdy, CWMcCurdy@lbl.gov
28. Nancy Meyer, NLMeyer@lbl.gov
29. Bill Nickles, nickless@mcs.anl.gov
30. Thomas Ndousse, tndousse@er.doe.gov
31. John Noe, jpnoe@sandia.gov
32. Bernard O'Lear, olear@ncar.ucar.edu
33. Juliet Pao, j.z.pao@larc.nasa.gov
34. James Patton, patton@cacr.caltech.edu
35. Don Petravick, petravick@fnal.gov
36. Walt Polanski, walt.polansky@science.doe.gov
37. Tom Potok, potokte@ornl.gov
38. Bill Rahe, whrahe@sandia.gov
39. Robb Ross, rross@mcs.anl.gov
40. Pat Schaefer, pschaef@us.ibm.com
41. Arie Shoshani, areie@lbl.gov
42. Alex Sim, ASim@lbl.gov
43. Horst Simon, HDSimon@lbl.gov
44. Ivan Sipos, Ivan.Sipos@digital.com
45. John Sobolewski, jssob@unm.edu
46. Danny Teaff, dannyt@us.ibm.com
47. Dick Watson, dwatson@llnl.gov
48. Thomas Wiggen, thomas_wiggen@und.nodak.edu
49. John Wilson, jdwilson@us.ibm.com
50. Dave Wiltzius, wiltzius@llnl.gov
51. Don Woelz, don@genroco.com