

# **PROBE PROJECT ACCOMPLISHMENTS**

**February 28, 2003**

**Prepared by  
Randall D. Burris  
Storage Systems Manager**

#### 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](mailto: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](mailto:reports@adonis.osti.gov)  
**Web site:** <http://www.osti.gov/contact.html>

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, 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 or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

**PROBE PROJECT ACCOMPLISHMENTS**

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 2003

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 .....	3
1. FINAL CONFIGURATION.....	3
2. SCIENTIFIC DISCOVERY THROUGH ADVANCED COMPUTATION .....	3
2.1 Scientific Data Management Integrated Software Infrastructure Center .....	3
2.1.1 Probe as “A Place To Be” .....	4
2.1.1.1 Hierarchical Resource Manager.....	4
2.1.1.2 Distributed Data Analysis.....	4
2.1.1.3 Linux Nodes.....	4
2.1.1.4 Agent Support.....	5
2.1.2 More Efficient Tertiary I/O .....	5
2.1.2.1 Selecting Data to Be Retrieved.....	5
2.1.2.2 Tape Support For PVFS.....	5
2.1.3 Improving Application Platform Independence .....	5
2.2 Earth Systems Grid II.....	6
2.3 DOE Science Grid.....	6
2.4 Terascale Supernova Initiative .....	6
2.5 Climate .....	6
3. DATA MINING AND DISTRIBUTED CLUSTER ANALYSIS .....	6
3.1 Mining Distributed Scientific Data from the Desktop.....	6
3.1.1 Distributed Cluster Analysis Algorithms .....	7
3.1.1.1 RACHET .....	7
3.1.1.2 Distributed Minimum Spanning Tree .....	8
3.1.2 Dimension Reduction .....	9
3.1.2.1 Distributed Principal Component Analysis.....	9
3.1.2.2 DfastMap .....	9
3.1.3 New Adaptive Data Reduction Algorithm .....	10
3.1.4 Summary .....	11
3.2 Computational Biology Research.....	12
3.2.1 Parallel Out-of-Core Enumeration of Extreme Metabolic Pathways .....	12
3.2.2 Hierarchical Feature Extraction.....	12
3.3 Computational Physics Research .....	13
4. NETWORK RESEARCH.....	13
4.1 Gigabit Ethernet.....	14
4.2 Web100.....	14
4.2.1 Webd .....	14
4.2.2 Work-Around Daemon.....	14
4.3 A TCP over UDP Test Harness.....	14
4.4 Optimizing Bulk Transfers in High-Delay/High-Bandwidth Networks.....	15

5.	DATA TRANSFER AND STORAGE DEVELOPMENT AND TESTING .....	15
5.1	Improve ORNL-NERSC bandwidth.....	15
5.2	HSI .....	15
5.2.1	Introduction .....	15
5.2.2	PROBE-Funded HSI Major New Features.....	16
5.2.2.1	Multiple Concurrent HPSS Connections (NDAPI Library).....	16
5.2.2.2	HSI Multiple Concurrent Connections.....	16
5.2.2.3	Striped Network Interfaces.....	16
5.2.2.4	Inter-HPSS “COPY” Feature .....	17
5.2.2.5	I/O Performance Improvements .....	17
5.2.2.6	Multi-Version Capability.....	17
5.2.2.7	Wide Area Network Transfer Improvements .....	17
5.2.2.8	Auto-Scheduling.....	18
5.2.2.9	Support for HSI on Windows Platforms.....	18
5.2.3	Other significant PROBE-Funded Enhancements.....	18
5.2.3.1	GLOBUS Support.....	18
5.2.3.2	Partial File Transfers.....	18
5.2.3.3	Support for HPSS.conf File .....	18
5.2.3.4	HSI Code Reorganization and Cleanup .....	18
5.3	Comparative Performance of HPSS Metadata-Management Alternatives.....	19
5.3.1	HPSS Metadata – Relational Database Testing.....	19
5.3.2	SFS Create/Delete Tests.....	19
5.4	Integrating HPSS with the GRID .....	19
5.5	HPSS Movers Using Gigabit Ethernet Network Connectivity and FibreChannel Disks .....	20
5.6	LTO-Associated HPSS Development .....	21
5.7	Remote Movers .....	21
5.8	Testing at the Request of the HPSS Collaboration.....	21
5.9	HPSSQ .....	22
5.10	HPSS Compatibility With New Infrastructure Products.....	22
5.11	High-Performance Visualization .....	22
5.12	Modeling Cache Performance in HPSS .....	22
5.13	Modeling Storage.....	23
5.14	Scheduled Transfer (ST).....	23
5.15	University of Vermont .....	23
5.16	Equipment Testing.....	24
5.16.1	Texas Memory Systems .....	24
5.16.2	SCSI-FibreChannel Bridge Testing.....	24
5.16.3	Lawrence Livermore National Laboratory (LLNL) S80 Test.....	24
5.16.4	Linear Tape Open (LTO) Test and Development .....	24
5.16.5	Gigabyte System Network (GSN) Hardware and Drivers.....	25
5.16.6	FibreChannel Tapes and Tape Striping with the StorageNet 6000 (SN0000).....	25
5.16.7	Storage Area Networks (SAN)/IP .....	26
5.16.8	Test of Gigabit Ethernet/FibreChannel Bridge.....	26
5.16.9	Linux Hosts .....	27
5.16.10	New Storage Devices .....	27

5.16.10.1 FibreChannel Disk Array .....	27
5.16.10.2 EMC Clariion .....	27
5.16.11 Low Cost Disk Arrays .....	28
5.16.12 iSCSI Tests .....	28
5.16.13 Two-Stripe T300-Gigabit Ethernet Jumbo-Frame Performance.....	29
5.16.14 StorageTek SCSI 9840 Tape Drive Testing.....	29
5.16.15 Configuration of StorageTek 9840 FibreChannel Tape Drives .....	29
5.16.16 Establishing an HPSS System on RS/6000 S80.....	29
5.16.17 Establishing an HPSS System on AIX 4.3.3 on an IBM H70.....	29
5.16.18 StorageTek FibreChannel Equipment Installation.....	30
6. SC2001 ACTIVITIES.....	30
6.1 Inter-HPSS File Transfer Demonstration .....	30
6.2 HPSS Wide-Area Remote-Mover Demonstration.....	30
6.3 Web-100-Tuned Wide-Area Bulk File Transfer Demonstration.....	30
6.4 Parallel Out-of-Core Enumeration of Extreme Metabolic Pathways Demonstration.....	31
6.5 Grid-Enabled PFTP Transfer Demonstration .....	31
7. PRODUCTION-RELATED PROJECTS.....	31
7.1 Projects Which Resulted in Production Implementation .....	31
7.2 Projects Which Discouraged Production Implementation .....	32
8. PUBLICATIONS AND PRESENTATIONS .....	33
8.1 Publications .....	33
8.1.1 Refereed Journal Papers .....	33
8.1.2 Refereed Conference Proceedings.....	33
8.1.3 Other Publications .....	34
8.1.4 Submitted/Written Papers.....	34
8.2 Patents.....	34
8.2 Presentations, Posters and Demonstrations .....	34
9. SUMMARY .....	36

APPENDICES

- APPENDIX A. ORNL EQUIPMENT DESCRIPTIONS
- APPENDIX B. NERSC EQUIPMENT DESCRIPTIONS



## LIST OF FIGURES

<b>Figure</b>		<b>Page</b>
1	Comparison of subsampling with adaptive data reduction over the course of a supernova simulation.....	10
2	Spherical Symmetry Instability Ranges Conserved under PC Compression.....	11
3	Grid Configuration.....	21
4	I/O Performance on EMC Clariion.....	27
5	Condor Performance.....	38



## INTRODUCTION

The Probe project 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 investigated local-area and wide-area distributed storage issues ranging from data mining to optimizing retrieval operations from tape devices.

Probe has completed its final year of operation. In this document we will describe the project accomplishments through September 30, 2002. 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 activities. The final sections will summarize production-related activities and publications.

Individual projects described in this document have used some Probe resources – 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 was funded by the Mathematical, Information, and Computer Sciences (MICS) department of the Advanced Scientific Computing Research office, Office of Science, and the U.S. Department of Energy (DOE).

## 1. FINAL CONFIGURATION

Appendices A and B list the equipment in Probe installations at ORNL and NERSC. Some of the equipment is heavily used for data-mining research. Other items support efforts to find and harden a production bulk-data transfer mechanism and other challenging storage and networking projects including, especially, understanding and optimizing Cray external I/O.

## 2. SCIENTIFIC DISCOVERY THROUGH ADVANCED COMPUTATION (SciDAC)

Probe facilities at ORNL were, and continue to be, used in several SciDAC projects.

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

Probe resources and staff have two roles in this project. First, Probe is a “place to be” – a testbed in which other elements of the ISIC are implemented and tested. The second role involves research and development into more efficient tertiary I/O.

### **2.1.1 Probe as “A Place To Be”**

Probe resources will continue to provide a prototyping environment for the use of other projects within the ISIC and the ISIC’s support for SciDAC Applications (in particular Terascale Supernova Initiative, Climate and High-Energy Nuclear Physics). A total of eight terabytes of fibrechannel Redundant Array of Independent Disks (RAID) disk capacity, five individual Linux nodes and two four-node clusters were procured for the use of these and other projects, to be applied to individual activities as necessary.

#### **2.1.1.1 Hierarchical Resource Manager**

ORNL/Probe provides two nodes, a Sun machine and an IBM machine, in support of two Grid projects (see 2.2 and 2.3). As part of the SDM-ISIC and the Earth System Grid II projects, NERSC researchers installed their Hierarchical Resource Manager (HRM). To support the HRM work, 280 gigabytes of fibrechannel disk capacity were attached to the Sun to act as a transfer cache. The Earth System Grid II activity performed experimental transfers of several hundred gigabytes of climate data between ORNL’s production High Performance Storage System (HPSS) and the National Center for Atmospheric Research; that capability is now in routine use. It is important to note that this transfer involves HPSS and non-HPSS installations.

#### **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 residing at different sites – distributed data access and analysis. Typical data mining applications require that the entire dataset exist at one site, on one machine. Such centralization is impossible with the massive data coming from high-energy physics, human genome, climate, etc. Consequently, research into mechanisms by which data can be analyzed without bringing all data to a single node has become important.

Prior to late FY2002 two ORNL/Probe RS/6000 nodes supported that type of research. The larger node, a six-processor S80, has been replaced with newer Power-4-based processors. This activity also makes heavy use of the “alice” cluster and a new Dell 6650 (see next section).

#### **2.1.1.3 Linux Nodes**

In FY2002 ORNL/Probe acquired four dual-processor Xeon nodes for use by SDM-ISIC researchers. The four nodes have now been formed into a cluster (the “Alice” cluster) being used by data research staff.

The following Linux resources have also been added:

- The “farkle” cluster, consisting of four identical nodes with a total of two terabytes of disk capacity, is being used to prototype a connection between HPSS and PVFS (Parallel Virtual File System) (see 2.1.2.2).
- Four Dell 2650 nodes (each with two 2.4 GHz Xeon processors, 2 gigabytes of memory, fibrechannel interfaces and 360 gigabytes of RAID disk capacity) are being used to support HRM, J-Lab’s Storage Resource Manager, Logistical Networking (see Section 2.4) and (soon) to characterize Cray I/O performance.

- One Dell 6650 node (four Xeon processors, 8 gigabytes of memory, 360 gigabytes of RAID disk capacity, fibrechannel interface) that is now in heavy use for data analysis requiring considerable compute power, memory and storage.
- One Consensys dual-processor node with two terabytes of disk capacity. This machine has been certified for use as an HPSS mover and as a source for HPSS NFS-based archiving operations. (It's interesting to note that ORNL testing preceded and influenced development of these capabilities in a collaboration between Consensys and IBM.) Both capabilities will be studied to evaluate performance. It is also to be used to investigate HPSS UNIX file I/O.

#### **2.1.1.4 Agent Support**

Agent technology is an important 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**

##### **2.1.2.1 Selecting Data To Be Retrieved.**

A typical application retrieves an entire file, and 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.

Improvements to the Hierarchical storage Interface (HSI) application (Section 5.2) have addressed this goal. HSI is now able to retrieve a group of chunks of a file specified in a list of <offset, length> tuples. Other new efficiencies in HSI are also discussed in Section 5.2.

##### **2.1.2.2 Tape Support For PVFS.**

PVFS was designed as a local, very fast parallel disk file system to support Linux clusters. It was not intended to support tertiary storage. However, using two new features of HPSS, such a link may be possible. To investigate the concept, ORNL acquired a duplicate of an Argonne cluster (the farkle cluster – see Section 2.1.1.3) and installed PVFS. Using HPSS's Linux data-mover capability and the ability to do I/O directly to UNIX filesystems (which underlie PVFS), we have prototyped a PVFS-HPSS link. Subsequent work will investigate performance optimizations and mechanisms for automatic migration of files from PVFS to HPSS.

#### **2.1.3 Improving Application Platform Independence**

Scientific applications usually run on a single platform, such as IBM/AIX or Linux.

Two data formats are in common use in various scientific communities – “NetCDF” and “HDF5”. Applications which use parallel I/O (via the MPI-IO mechanism) have been limited to using HDF5; no parallel implementation of NetCDF existed. One SDM-ISIC project integrated the NetCDF application

programming interface with the “ROMIO” implementation of MPI-IO and subsequently testing with IBM’s MPI-IO and HPSS/MPI-IO. The success of that test indicates that an application using MPI-IO can use either the HDF5 or the NetCDF data structure. Such an application can also run over PVFS (which can underlie ROMIO) or over HPSS, leading to much improved application platform independence. ORNL installed the HPSS/MPI-IO software and acquired the farkle cluster for native PVFS, providing a testbed for applications using either data format.

## **2.2 EARTH SYSTEMS GRID II**

The use of Probe resources was specified in the ORNL portion of the Earth Systems Grid II proposal. To support it, ORNL/Probe supplied an RS/6000 Model 44P-170 running the AIX operating system and a Sun E250 running Solaris and augmented both nodes with 280 GB of SCSI disk capacity. The Sun is being used by SciDAC SDM ISIC staff working with the Hierarchical Resource Manager (see 2.1.1.1); the IBM node is used by Argonne staff for debugging AIX Globus 2.0 software.

## **2.3 DOE SCIENCE GRID**

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

## **2.4 TERASCALE SUPERNOVA INITIATIVE (TSI)**

This SciDAC Application requires significant storage resources, processing power and network bandwidth to support visualization of the massive datasets produced by its simulation codes. ORNL/Probe resources are being used as researchers determine how to select, render and transport data to visualization equipment – probably across wide area links and to multiple destinations.

In a separate initiative, Logistical Networking technology from the University of Tennessee (UT) is being considered to support some of the requirements of TSI. UT staff have created a depot on one of the Dell 2650 nodes mentioned in Section 2.1.1.3.

## **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 on Probe resources, as described in the next section. Finally, studies of file transfers via HRM from ORNL’s HPSS to the National Center for Atmospheric Research – a site which does not have HPSS – proved successful (see Section 2.1.1.1) and are now routine.

## **3. DATA MINING AND DISTRIBUTED CLUSTER ANALYSIS**

The projects described in this section made heavy use of several Probe nodes. The research continues; the researchers continue to make heavy use of individual and clustered Linux machines and one of the IBM RS/6000 nodes.

### 3.1 MINING DISTRIBUTED SCIENTIFIC DATA FROM THE DESKTOP

As Cluster Computing and the Grid are becoming the paradigms of current and future high-performance computing, massive petascale data sets distributed over a network of clusters or a Data Grid are the future in science and, particularly, simulation science. Even current massive data sets stored in multiple files, multiple disks, or multiple tapes are often too large for centralized processing. Some examples are medical records of distributed groups of individuals, sales records of distributed stores on the same group of products, climate simulations based on different initial and boundary conditions, genome characteristics for different organisms, etc. More applications are becoming available as data standards are developed across distributed locations. Wegman discusses the interaction of huge data sets and the limits of computational feasibility, concluding that most current data analysis techniques break down on data sets beyond about 10 gigabytes (analysis of which using current algorithms would require about three years on a teraflop machine). This is also true of cluster analysis.

ORNL's data mining effort within DOE's Probe project concentrated on developing methods for clustering scientific data distributed over a computational grid or on the Internet. The central concept for these methods is to perform local analyses on local data and combine the results into a global analysis with very low communication and data transfer requirements. The methods developed are known as "RACHET" to signify that the cluster information flow is one way from the massive distributed data sets to the global analysis results on the desktop of a scientist.

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 the 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 and transfers or communicates only minimum summary information to a merger site (e.g., desktop), where these summaries are combined into a global analysis.

Three major methodologies resulted from this work. The first methodology is focused on distributed cluster analysis algorithms. It includes a RACHET algorithm [Samatova et al., 2001.a and 2002.a] for merging local hierarchical cluster analyses that is effective for a variety of local clustering techniques and requires little communication. ORNL also developed an algorithm for computing an approximate minimum spanning tree of distributed data to provide another fast clustering capability in RACHET with a single linkage criterion [Samatova et al., 2003.a].

The second methodology targets distributed dimension reduction algorithms and coupling them with cluster analysis algorithms. This methodology can be viewed as an alternative method for reducing communication in RACHET cluster analysis and sometimes improving its efficacy. This category, includes two distributed dimension reduction algorithms: Distributed Principal Component Analysis

(DPCA) [Qu et al., 2002 and Ostrouchov et al., 2003.a] and Distributed FastMap (DFastMap) [Abu-Khzam et al., 2002 and Samatova et al., 2003.b].

Finally, the third methodology is focused on developing data reduction algorithms to decrease the amount of data that is being transferred over the network or that is being stored on disks/tapes. The new data reduction algorithm for simulation data that is based on Principal Component Analysis (PCA) of grid field blocks [Ostrouchov et al., 2003.b]. The algorithm outperforms subsampling by a factor of three while maintaining comparable mean squared error.

These new RACHET algorithms are being deployed and made available to simulation scientists through Adaptive Simulation Product Exploration and Control Toolkit (ASPECT) client-server software that will soon be released under the Scientific Data Management ISIC SciDAC project.

### **3.1.1 Distributed Cluster Analysis Algorithms**

#### **3.1.1.1 RACHET**

Rachet is a hierarchical clustering method for analyzing multi-dimensional distributed data [Samatova et al., 2001.a and 2002.a]. A typical clustering algorithm requires bringing all the data to a centralized warehouse. This results in  $O(nd)$  transmission cost, where  $n$  is the number of data points and  $d$  is the number of dimensions. For large datasets this is prohibitively expensive. In contrast, RACHET runs with at most  $O(n)$  time, space, and communication costs to build a global hierarchy of comparable clustering quality (see Figure 1) by merging locally generated clustering hierarchies. RACHET employs the encircling tactic in which the merges at each stage are chosen so as to minimize the volume of a covering hypersphere. For each cluster centroid, RACHET maintains descriptive statistics of constant complexity to enable these choices. RACHET's framework is applicable to a wide class of centroid-based hierarchical clustering algorithms, such as centroid, medoid, and Ward.

#### **3.1.1.2 Distributed Minimum Spanning Tree**

Most clustering algorithms are impractical when dealing with extremely large and high dimensional data sets. The problem gets harder when such data sets reside on a number of geographically dispersed machines. In this case, the communication cost is expected to dominate the execution time. Our main approach is based on the use of minimum spanning trees (MSTs). This well known technique is called Single Linkage clustering.

The initial centroid-based approach to cluster definition in RACHET has been extended to single linkage cluster definition by our new algorithm for computing an approximate MST for distributed data [Samatova et al., 2003.a]. The MST is the key component of single-linkage clustering algorithms and contains information sufficient for computing single-linkage clusters. Single linkage clustering has been very successful with non-Gaussian clusters. It also is a clustering algorithm with low computational complexity, thus being able to deal with massive data sets.

## 3.1.2 Dimension Reduction

### 3.1.2.1 Distributed Principal Component Analysis

Dimension reduction is a necessary step in the effective analysis of massive high-dimensional data sets. It may be the main objective in the analysis for visualization of the high-dimensional data or it may be an intermediate step that enables some other analysis such as clustering. Principal Component Analysis (PCA) (also known as the Karhunen-Loeve procedure, eigenvector analysis, and empirical orthogonal functions) is probably the oldest and certainly the most popular technique for computing lower-dimensional representations of multivariate data. The technique is linear in the sense that the components are linear combinations of the original variables (features), but non-linearity in the data is preserved for effective visualization. The technique can be presented as an iterative computation of the direction of highest variation followed by projection onto the perpendicular hyperplane. This technique quickly provides a few perpendicular directions that account for the majority of the variation in the data, giving a low dimensional representation of the data. A complete set of principal components can be viewed as a rotation in the original variable space.

ORNL developed a new method for computing a global principal component analysis for the purpose of dimension reduction in data distributed across several locations [Qu et al., 2002 and Ostrouchov et al., 2003.a]. It assumes that a virtual  $n \times p$  (items  $\times$  features) data matrix is distributed by blocks of rows (items), where  $n > p$  and the distribution among  $s$  locations is determined by a given application. The approach is to perform local PCA on local data without any data movement and then move and merge the local PCA results into a global PCA. The representation of local data by a few local principal components greatly reduces data transfers with minimal degradation in accuracy.

Most high-dimensional data have lower intrinsic dimensionality thus allowing a good lower-dimensional representation. Existing methods that bring data to a central location require  $O(np)$  data transfer even if only a few principal components are needed. In the worst case, when an exact PCA is computed, the new algorithm is  $\min(O(np), O(sp^2))$ . It is of  $O(sp)$  data transfer complexity when intrinsic dimensionality is low or when an approximate solution is sufficient. The ability to vary data transfers by controlling precision provides a great deal of flexibility.

Incorporating the distributed principal components algorithm into the RACHET cluster analysis framework will improve clustering performance and provide a control parameter that can vary the amount of approximation needed for available network bandwidth. The concepts developed for distributed data can also be used to design updating methodology for clustering to deal efficiently with dynamically growing data sets.

### 3.1.2.2 DFastMap

It is well known that information retrieval, clustering and visualization can often be improved by reducing the dimensionality of high dimensional data. Classical techniques offer optimality but are much too slow for extremely large databases. The problem becomes harder yet when data are distributed across geographically dispersed machines. To address this need, an effective *linear time* distributed dimension reduction algorithm, DFastMap, was developed [Abu-Khzam et. al., 2002 and Samatova et al., 2003.b].

DFastMap algorithms map high dimensional objects distributed across geographically dispersed machines into points in lower dimensional space, so that distances between the objects in data space are preserved as much as possible. Transferring all local data to a central location and running the centralized version of this algorithm would require  $O(nd)$  data transmission, where  $n$  is the number of objects and  $d$  is the number of features. DFastMap algorithms require only  $O(kst)$  data transmission, where  $s$  is the number of data locations and  $k$  is the dimensionality of the projected space. Thus, it runs in linear time and requires very little data transmission.

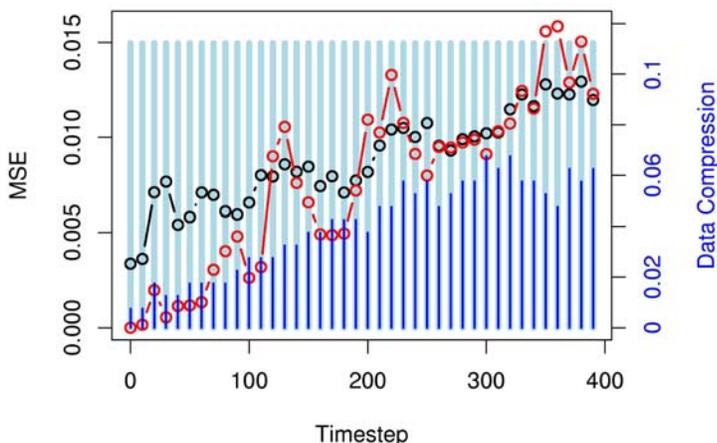
A series of experiments were conducted to gauge how the algorithm’s emphasis on minimal data transmission affects solution quality. Stress function measurements indicate that the distributed algorithm is highly competitive with its centralized version.

### 3.1.3 New Adaptive Data Reduction Algorithm

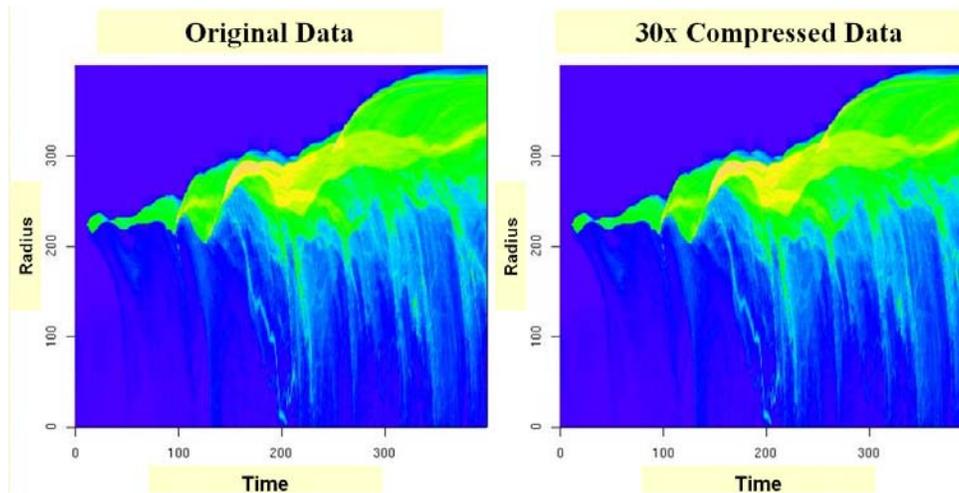
Many scientific problems that are currently explored by simulation, such as those in climate and astrophysics, routinely produce data sets of many gigabytes and often terabytes. Analysis and visualization of such data sets is usually cumbersome and is limited by the need to transmit the data to a scientific workstation.

A data reduction methodology developed at ORNL addresses the data transmission bottleneck for simulation data that is a time series of two or three-dimensional fields [Ostrouchov et al., 2003.b]. An application of the Adaptive Data Reduction technique to data from the SciDAC Terascale Supernova Simulation project produced reduction factors ranging from 20 to 200 over the course of the simulation (see Figures 2 and 3). This reduction dramatically increased the scientists’ ability to understand the simulation results. When compared to subsampling, data transmission was reduced by a factor of three while matching mean squared error of approximation.

The new method only transmits a reduced data representation for each time period, which is then used to reconstruct the original field to a pre-specified precision. The field is partitioned into spatially contiguous segments, thus exploiting spatial homogeneity in small areas. A singular value decomposition based algorithm that computes a minimal set of pseudo-segments, which are linear combinations of actual segments, exploits similarities between segments. The resulting set of pseudo-segments and a set of



**Figure 1.** Comparison of subsampling with adaptive data reduction over the course of a supernova simulation. Lines with symbols show mean squared error (black is subsampling and red is adaptive) and vertical bars show proportion of data transmitted (light is subsampling and dark is adaptive).



**Figure 2.** Spherical Symmetry Instability Ranges Conserved under PC Compression.

weights are used to optimally reconstruct the field for a specified level of precision. The amount of data required adapts to the amount of similarity among groups of segments.

### 3.1.4 Summary

New data mining and data reduction algorithms developed under this project are being deployed within the ASPECT task in the SciDAC Scientific Data Management ISIC (PI: Arie Shoshani, LBL). They will enable terascale analysis of distributed and dynamically changing scientific datasets. Continuing work to implement package, and make the algorithm robust will proceed 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 also be leveraged under the National Center for Supercomputing Applications TeraGrid initiative in FY2003 with application to computational biology research performed in collaboration with Argonne National Laboratory (PI: Natalia Maltsev).

## 3.2 COMPUTATIONAL BIOLOGY RESEARCH

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 the Genetic Circuit Research Group of the University of California at San Diego (UCSD) (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% [Samatova et. al., 2002.b].

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 door open for further improvements.

### **3.2.1 Parallel Out-of-Core Enumeration of Extreme Metabolic Pathways**

A highly scalable algorithm that enumerates all the extreme metabolic pathways on a genome-scale was demonstrated on a cluster of Linux PCs [Samatova et. al., 2003.c]. The generation process for a selected example took on the order of fifteen minutes on four processors by the new algorithm as opposed to five days by the original code obtained from the UCSD. Each iteration showed 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 will then be used to analyze, interpret, and predict metabolic functioning and control of a living cell.

### **3.2.2 Hierarchical Feature Extraction**

High throughput genome sequencing projects have been producing completely sequenced genomes at an accelerated rate, but the functions of a substantial portion of the predicted open reading frames (ORFs) are unknown. The number of functionally uncharacterized ORFs continues to rise. All this functionally uncharacterized sequenced data begs for methods that are able to rapidly and accurately predict the function of individual gene products, or proteins, from their sequence through comparison with other known genes and genomes. One of the main principles underlying current methods for protein function characterization is the correlation of function with the sequence similarity or three-dimensional structure similarity, by which functional information is transferred from known proteins to the unknown proteins (the so-called "guilty-by-association" principle). However, due to a number of factors, these transfers can sometimes be extremely uncertain resulting in inaccurate function inferences.

To address these limitations, we developed a method of hierarchical multi-resolution functional classification of proteins [Samatova et al., 2003.d]. The taxonomy of protein functional groups is created automatically with manual curation using proteins with known functions from the Swiss-Prot. The highly conserved (high resolution) protein functional groups are at the leaves of the hierarchy. The method is based on extracting discriminating features that are specific to each node of the hierarchy and applying a machine learning method to classify its children nodes. The best method is selected based on performance results for several discriminative methods, such as support vector machines, association rules, and decision trees. This method results in highly-conserved functional group whose functions are clearly different from one another. The method is also scalable with the number of functional groups since at each node only the locally computed features are considered. At the same time, it is the specificity of the locally extracted features that is the main driver behind the method's accuracy. This method predicts functional roles for uncharacterized *Synechococcus sp* ORFs, a result which is very important for the DOE Genomes To Life Project (ORNL PI: Al Geist).

### 3.3 COMPUTATIONAL PHYSICS RESEARCH

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 ( $\sqrt{3}\times\sqrt{3} \leftrightarrow 3\times 3$ ) transition and a disorder-order transition in the defect distribution.

Two phenomenological models have been developed that describe these transitions and their interrelation [Melechko et. al, 2001]. 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  $3\times 3$  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.

### 3.4 PARTICLE PHYSICS SUPPORT

Over the past four years, solar neutrino experiments at the Super-Kamiokande Observatory in Japan and the Sudbury Neutrino Observatory (SNO) in Canada have offered compelling evidence that neutrinos have nonzero mass and that the electron, muon and tau neutrinos are actually different states of the same particle. While the evidence was strong, supportive evidence involving terrestrial anti-neutrinos could rule out one of the sources of possible error.

Japan's KamLAND, the Kamioka Liquid scintillator Anti-Neutrino Detector, was used to do the investigation. Supported by an international collaboration (including DOE), KamLAND is the largest low-energy anti-neutrino detector ever built. KamLAND's 1,879 photomultiplier sensors detect flashes of light resulting from the collision of an anti-neutrino (generated by nuclear reactors in Japan and Korea) with a proton; the resulting pulses are converted to signals which are recorded for later analysis.

KamLAND experiments began generating about 200 gigabytes of data per day in January 2002; the data were stored on LTO tape cartridges. After six months, U.S. scientists running experiments at KamLAND had 800 tapes containing a vast amount of data, and they wanted to present initial results at conferences in September.

Fortunately, the NERSC Center had an LTO system on loan from IBM for the Probe storage research project (see Sections 5.6 and 5.16.4). IBM agreed to extend the loan for a few months so the KamLAND group could transfer their data to HPSS. The data transfers were done at night and on weekends, increasing the HPSS data traffic by up to 80 percent per day. In all, more than 48 terabytes of data were transferred from the tapes to HPSS. The KamLAND group then used the Parallel Distributed Systems Facility to analyze their data, sometimes using the full 400-processor cluster.

The results were presented at the International Workshop on Neutrinos and Subterranean Science in Washington, DC, September 19-21, 2002, and the 16th International Conference on Particles and Nuclei

in Osaka, Japan, September 30 through October 4, 2002. The results were also submitted for publication in Physical Review Letters.

## **4. NETWORK RESEARCH**

Network research has been an important part of Probe activity. In this section we will briefly describe various projects undertaken by Tom Dunigan and Florence Fowler of ORNL. A very extensive and informative set of Web pages developed by Tom Dunigan describes 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 GIGABIT ETHERNET**

An early project in Probe at ORNL was to verify the compatibility of HPSS and supercomputers with jumbo-frame Gigabit Ethernet networks. We found no incompatibilities, and performance evaluations showed considerable improvements in throughput and a reduction in CPU usage. The entire ORNL supercomputing and HPSS network is now based on jumbo-frame Gigabit Ethernet.

### **4.2 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>.

#### **4.2.1 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 text file suitable for statistical analysis or auto-tuning. The statistics will feed a database to be developed in the Net100 project.

#### **4.2.2 Work-Around Daemon**

ORNL 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 six distant institutions and over three network technologies. Wide-area networks included ESnet (OC12/OC3), UT (BR/OC3) and Internet 2. Local-area networks included 100-base T and Gigabit Ethernet (including jumbo frames). See <http://www.csm.ornl.gov/~dunigan/netperf/wad.html>.

A production version of the WAD is under development.

### **4.3 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 acknowledgement, optional delayed acknowledgement, sliding window, timeout-retransmissions with rate-based restart, bigger initial window, bigger maximum segment size, burst avoidance, congestion avoidance (are more aggressive, experimenting with initial window size and “additive increase multiplicative decrease” parameters).

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

### **4.4 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. Because our sites are so far apart, latency is high (>60 ms round-trip time); TCP’s congestion avoidance has been shown to 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**

When we began, the effective 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 have allowed. An initial project to find and remedy the cause was completed in FY2000. 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, became more commonplace. 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, has provided the following final report on Probe-funded improvements and enhancements to HSI.

### **5.2.1 Introduction**

The HSI is a command line tool that is extensively used by the HPSS community. It provides a UNIX-like user-friendly environment that can be used in both interactive and batch mode, and runs on all major UNIX-based platforms. Its many features include recursion for most commands, e.g. listing, storing or retrieving an entire tree of files with a single command, and a variety of commands to simplify system administration.

The HSI Non-DCE Client API Library (NDAPI), Architecture-Independent Threads Library (AI\_THREADS) and API Extensions Library (API\_EXTENSIONS) collectively comprise a separate standalone framework to allow applications to run on non-DCE platforms. The HSI NDAPI library code was used as the basis for the HPSS non-DCE Gateway, which provides a subset of the capabilities of the HSI version.

Probe has funded a number of major new capabilities that exist in the current version of HSI and its component libraries, as well as improvements to preexisting features. The remainder of this report provides a description of these enhancements.

Probe funding of HSI has significantly contributed to the success of HPSS at sites around the planet. Most sites now use it as a tool for their internal system administration and/or provide it to their user community. At some sites it has replaced the use of Network File System and Distributed File System interfaces to HPSS, and many sites have repeatedly requested that HSI become a part of the HPSS product. Among the HSI production sites are 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.

### **5.2.2 Probe-Funded HSI Major New Features**

#### **5.2.2.1 Multiple Concurrent HPSS Connections**

This feature adds API interfaces which allow HPSS Client API programs (HSI as well as others) to open and close multiple concurrent connections to the same or different HPSS systems. The NDAPI library was multithreaded as a part of this project, as well as adding code to manage problem areas such as

duplicated file descriptors when working with multiple HPSS systems. The connection management code was augmented to support different authentication mechanisms for each connection.

### **5.2.2.2 The “logical drive”**

This feature was added, making use of the new NDAPI library Multiple HPSS APIs, to provide the ability for users to open connections to multiple HPSS systems without requiring DCE cross-cell trust, and without requiring HPSS junctions to be created in order to give the appearance of a single namespace. Instead, the simple but powerful "logical drive" concept was introduced, so that each HPSS connection appears to the user as a different disk drive; file pathnames make use of an optional "drive letter" prefix (e.g. "A:somepath/somefile") to reference files associated with a particular connection. All HSI commands that reference HPSS files were modified to support the optional drive letter prefix.

### **5.2.2.3 Striped Network Interfaces**

This feature provides improved performance by making use of multiple network interfaces on the HSI client host, based upon a configuration file on the client host.

### **5.2.2.4 Inter-HPSS "COPY" Feature**

This feature makes use of the HSI multiple connection features for the case when files are being copied between different HPSS systems, as determined by the logical drive letters used to prefix the source and sink pathnames. Three methods of accomplishing inter-HPSS file transfers were experimentally implemented. Two were kept in the final version:

- "local" method, in which HSI reads from one system and writes to the other, so that it is in the middle of the data transfer.
- "server" method, which provides for 3rd party copies between HPSS systems. In this method, HSI controls the transfer, but the data are passed directly from the movers of the source HPSS to the NDAPI server on the sink HPSS system.

A command option provides the user with the ability to control the method selection; the default is to use the "server" method.

### **5.2.2.5 I/O Performance Improvements**

Probe supported work adding a "buffer pool" concept, which essentially provides double-buffered I/O for multiple threads, independently of the HPSS striping. It should be noted that the HSI I/O and striped network mechanisms were used as the basis for implementation of the “HTAR” (HPSS and TAR) utility at Lawrence Livermore National Laboratory (LLNL). The success of HTAR at LLNL and the reduced development time for its implementation are a direct result of the Probe-funded HSI I/O improvements.

### **5.2.2.6 Multi-Version Capability**

Some data structures used by the HPSS Client API are changed for different versions of HPSS. The usual practice of building a separate version of client API programs for use with each HPSS version does not work for HSI, since a single executable must be able to communicate with multiple HPSS systems, each

of which may be at a different HPSS level. In order to solve this problem, protocol messages to determine the NDAPI Server's HPSS version were added, and a dynamic translation layer was added to the NDAPI library, which transparently converts data structures from the form used on the NDAPI Server's HPSS system to the form expected by the client program. For further information see <http://www.csm.ornl.gov/PROBE/hsi.html>.

### **5.2.2.7 Wide Area Network Transfer Improvements**

In order to improve the network transfer performance when copying files between HPSS systems, the optional use of multiple network sockets may now be specified when using the *server* method (see 5.2.2.4). In particular, this use of multiple sockets can help to compensate for the TCP slow restart congestion control algorithms that are invoked when network saturation occurs.

For the *local* inter-HPSS copy method, HSI and the NDAPI library were modified to add "extended I/O" API calls, which provide the ability to use a private socket connection for data transfers instead of using the HPSS mover protocol; use of this capability significantly reduces the number of protocol messages that are required for each data block. This reduction is particularly important for inter-HPSS copies across a wide-area network, due to the latency resulting from sending control messages at the beginning and end of each data block.

### **5.2.2.8 Auto-Scheduling**

Several HSI commands may cause tape mounts to occur. For example, when fetching files from HPSS, it may be necessary to mount a tape in order to stage the file onto disk cache. Since tape mounts and positioning are costly operations in terms of time and resource use, it is important to optimize the commands that cause these operations to occur, so as to minimize the number of tape mounts, and, to a lesser extent, to minimize the number of positioning operations that take place when several files are fetched from the same tape. A common problem that occurs when fetching files that live on different tapes is to mount the first tape, read a file, unload the tape, mount a second tape, read a file, unload the tape, and then remount the first tape again to fetch another file.

Probe has funded development of "auto-scheduling" Extension Library APIs, and changes to HSI to make use of these new APIs, for commands that may cause tape mounts (e.g., "stage", "chcos", "get"). The new APIs provide the ability to optimize tape mounts and seeks within tapes, and to overlap background staging of files that live on tape with retrievals of files that are on disk cache.

### **5.2.2.9 Support for HSI on Windows Platforms**

A significant percentage of the user community uses Microsoft Windows platforms to do their work. HSI was developed for use on UNIX-based operating systems, and thus was unusable on Windows machines. In order to provide HSI for Windows users, HSI was ported to the Cygwin environment. Cygwin (<http://www.cygwin.com>) is freely available for non-commercial use. As part of this port, Globus GSI authentication was also debugged and tested on Windows under Cygwin.

### **5.2.3 Other Significant Probe-Funded Enhancements**

Other significant Probe-funded improvements to HSI and its supporting libraries include the items below.

#### **5.2.3.1 GLOBUS Support**

The NDAPI library was modified to support Globus GSI Authentication as a means of establishing credentials for an HSI session, and HSI was modified to add this as an additional optional authentication mechanism. GSI authentication for both Globus 1 and Globus 2 (once it became available) were tested; Globus 2 support was added to the baseline HSI version.

#### **5.2.3.2 Partial File Transfers**

In order to potentially reduce the amount of data that is transferred across a WAN, HSI was modified to support partial file retrievals, using a list of user-specified offsets and lengths. In addition, options were added to the "put" and "get" commands to allow restarting of previously failed transfers from the point of failure, rather than having to retransmit data, which already had been transferred successfully.

#### **5.2.3.3 Support for HPSS.conf File**

HPSS 4.3 consolidated several configuration files into a single "HPSS.conf" file. A set of APIs was added to the API Extensions Library to facilitate the use of this file for lookup of network options, etc.

#### **5.2.3.4 Recursive Inter/Intra File Copy Capability**

The ability to perform recursive intra-HPSS and inter-HPSS copies (using the "cp" command) was added. Previously the command could only be used in non-recursive mode.

#### **5.2.3.5 Intra-HPSS Fast File Copy Capability**

HPSS version 4.3 provides an interface to the Bitfile Server to copy an HPSS file internally without requiring the data to be transferred to and from client buffers, subject to the restriction that the source and sink files live in the same storage subsystem. HSI and the NDAPI library were modified to add the Client API call to support this function, and to automatically make use of the new API within HSI, if possible, for the "cp" command.

#### **5.2.3.6 Namespace Object Annotation Capability**

HSI was modified to add an "annotation" option for commands that create HPSS files and directories. A new "annotate" command was added to provide the ability to add/change/remove annotation for existing files and directories, and a new listing option was added to display the annotation.

## **5.3 COMPARATIVE PERFORMANCE OF HPSS METADATA MANAGEMENT ALTERNATIVES**

### **5.3.1 HPSS Metadata - Relational Database Testing**

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>.

### **5.3.2 SFS Create/Delete Tests**

LANL tested two RS/6000 servers with different memory architectures (switched vs. bus). They found that Encina/SFS, the repository for HPSS metadata, exhibited exceptionally good performance on the bus machine. The Model S80 at ORNL has the switched memory architecture, so we investigated the S80's performance on the same tasks. For metadata-only tasks, the performance of the S80 compared well to the bus architecture results at LANL. See <http://www.csm.ornl.gov/PROBE/sfs.html> for more information.

## **5.4 INTEGRATING HPSS WITH THE GRID**

NERSC committed to support Grid-enabled applications and used the NERSC/Probe environment to develop, evaluate, and test various Grid technologies. NERSC's File Storage Group worked with other national laboratories and universities to better integrate Grid applications and HPSS.

NERSC is part of the working group involving the Argonne, Los Alamos, Sandia and Lawrence Berkeley National Laboratories (ANL, LANL, SNL, LBNL) and IBM to evaluate, develop and test the integration of GridFTP with HPSS. One project added Grid extensions to the HPSS PFTP server to handle grid enabled FTP applications; a second extended work on that server to address various client idiosyncrasies. Most GridFTP clients (e.g., globus-url-copy, gsyncftp, gsi enabled PFTP) are now supported (see Figure 4).

Probe supported the Grid enabled HSI interface to HPSS (see Section 5.2.3.1). Users with Grid certificates can use HSI to transfer files to and from HPSS. NERSC staff is working with the Grid Technologies Group at LBNL to provide a Grid Web Services portal into HPSS.

One of the results of this collaboration, a modification of the Globus File Yanker, was demonstrated at SC2002. Globus File Yanker provides a convenient, light-weight Web interface to filesystems served by GridFTP and Grid-enabled HPSS FTP servers, and provides the ability to directly perform a third party

copy between these servers. The user can log into the system from a standard Web browser. See <http://hpcf.nersc.gov/storage/hpss/probe/gfy/gfy.pdf> for a technical description.

These capabilities, and others from other NERSC projects, will be integrated into the NERSC production HPSS archival data storage system to support the DOE Science Grid.

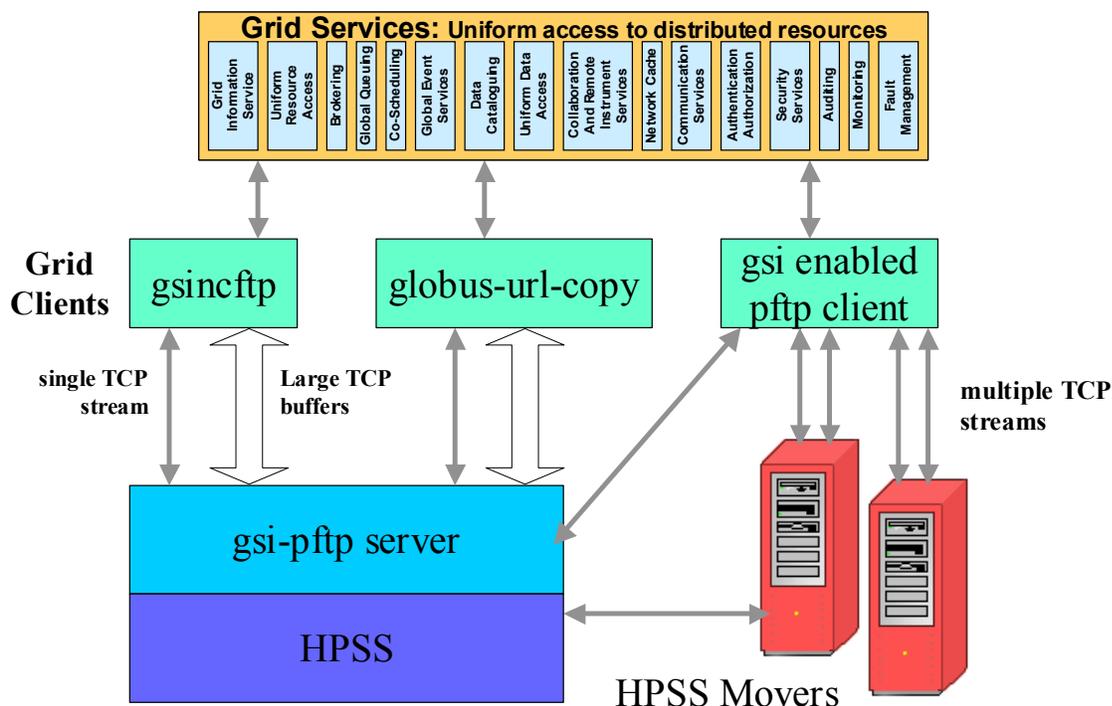


Fig. 3. Grid Configuration

### 5.5 HPSS MOVERS USING GIGABIT ETHERNET NETWORK CONNECTIVITY AND FIBRECHANNEL DISKS

ORNL purchased servers from IBM, Compaq, Silicon Graphics and Sun with the goal of testing/tuning HPSS mover operation using FibreChannel disks and Gigabit Ethernet network interfaces, neither of which had been included in HPSS testing anywhere. 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 Storage System Management and the mover. That capability was released in HPSS version 4.3.

This work is described at [http://hpcf.nersec.gov/storage/hpss/probe/LTO/IBM\\_LTO\\_test\\_1.pdf](http://hpcf.nersec.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 shown the concept is viable. When the two production HPSS installations are at the same HPSS release level, we will consider production deployment.

## **5.8 TESTING AT THE REQUEST OF THE HPSS COLLABORATION**

Several projects assisted HPSS support and testing. 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. Later, ORNL became a beta-test site for HPSS 4.2 for the purpose of testing HPSS Solaris core servers and movers in a mixed IBM/Solaris environment.

Also, ORNL/Probe equipment was used to test HPSS version 4.3 on AIX and Solaris with StorageTek 9840 tape drives. IBM staff performed the tests. In each case the testing was successful and the product has been released. See <http://www.csm.ornl.gov/PROBE/P2projects.html>.

## **5.9 HPSS QUEUEING (HPSSQ)**

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 HPSSQ command, either interactively or in batch, which will be communicated to and executed by a separate server. One benefit is to free the user from waiting for a transfer to take place. A second benefit is that transfers will not require that HPSS be available at the time the request is made. This capability 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 with a specific set of infrastructure products (including DCE, Encina, Encina's Structured File System and Sammi) and on two platforms, IBM/AIX and Sun/Solaris. The various products have different release schedules; frequently, an infrastructure product comes out with a new release shortly after an HPSS 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.

Throughout its operation ORNL/Probe installed the latest release of HPSS over the latest releases of infrastructure products. This task involved compiling, building and running HPSS in the new environment; combinations included HPSS 4.2 over DCE 3.1 and HPSS 4.3 and HPSS 4.5 over AIX 5.1. The project provided HPSS customers with confidence that HPSS operated correctly with the latest infrastructure.

## **5.11 HIGH-PERFORMANCE VISUALIZATION**

ORNL studied Gigabyte System Network (GSN) to evaluate its performance in a heterogeneous environment (i.e., with IBM or Compaq supercomputers feeding visualization data to an Origin 2000). The GSN testing demonstrated fairly good performance – setting some records that were reported in the trade press – but at too great an expense per port. At the time of the testing, furthermore, no PCI-bus computers were capable of transferring data as fast as GSN could handle it, so performance could not improve. Consequently, GSN was not useful in the ORNL environment and the GSN equipment has been removed. 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.

The study provided data on file size distributions and on access patterns. A cache simulation was performed which provided data on cache hit ratios and consequent tape mounts needed to retrieve requested files. Minimum, maximum and average cache residence times were extracted from the simulations. Sensitivity analyses were performed for the number of tape mounts required as a function of the size of cache, for the number of tape mounts required as a function of the disk allocation size, and for the number of tape mounts required as a function of the purge policy choices.

See <http://hpcf.nersc.gov/storage/hpss/probe/caching/cache-behavior.pdf> "Exploration of Cache Behavior Using HPSS Per-File Transfer Logs" for details.

## **5.13 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 initiated 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) were to be added later as additional projects.

“OPNET”, a network-modeling tool, was purchased and installed. Discussions with simulating staff from IBM and StorageTek were held; various people were willing to participate. Early on, students at the University of North Dakota and their advisor undertook elements of the simulation, but unfortunately were not able to see the project through to completion. Other attempts to find staff to develop the model eventually failed as well.

#### **5.14 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.

Our subsequent research dimmed hopes of making effective use of Scheduled Transfer between heterogeneous nodes. One problem was that the specification had been implemented only on SGI Origin equipment. Second, the protocol is very light-weight, with very little error correction, so it is not appropriate for other than local-area networks. Third, the code is difficult to implement.

Clear and significant performance gains would be necessary to justify the effort to develop ST applications, given the problems stated. Studies have shown only minimal throughput gains for ST when compared with Gigabit Ethernet jumbo frames. ST work was discontinued.

#### **5.15 UNIVERSITY OF VERMONT**

A researcher at the University of Vermont had a need for a data set of significant size for use in a data-mining project. Probe supplied a 5 GB ORNL global climate dataset. See <http://www.csm.ornl.gov/PROBE/bigdata.html>.

#### **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.

ORNL tested raw performance, verified RAM-SAN compatibility with HPSS, re-ran the database testing described in Section 5.3 and characterized performance using a RAM-SAN as a rotating disk mirror. See [http://www.csm.ornl.gov/PROBE/TMS\\_ORNL.html](http://www.csm.ornl.gov/PROBE/TMS_ORNL.html) for more information.

NERSC ran three different benchmarks. Initial baseline timing benchmarks used the UNIX utility dd. To benchmark transactional performance NERSC used the Encina database system 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 purchased the device for production HPSS 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 Lawrence Livermore National Laboratory (LLNL) S80 Test**

The first project completed in ORNL's Probe installation was a study of the CPU performance of the IBM RS/6000 Model S80 Enterprise Server, performed at the urgent request of Lawrence Livermore National Laboratory. LLNL had a tight deadline that required considerable effort to achieve, but the testing was concluded on time. Two letters of appreciation resulted. Summarizing the results, the S80 demonstrated excellent CPU performance and the ability to sustain very high single-channel and aggregate I/O throughput. See <http://www.csm.ornl.gov/PROBE/S80.html> for details.

### **5.16.4 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 was completed in FY2001; the capability was released in HPSS 4.3. It is described at [http://hpcf.nersc.gov/storage/hpss/probe/LTO/IBM\\_LTO\\_test\\_1.pdf](http://hpcf.nersc.gov/storage/hpss/probe/LTO/IBM_LTO_test_1.pdf).

### **5.16.5 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 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, expecting 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 available to 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.

#### **5.16.6 FibreChannel Tapes and Tape Striping with the StorageNet 6000 (SN6000)**

Both sites tested the management of FibreChannel tape drives using a StorageTek SN6000. The SN6000 Probe hosts and the STK 9940 FibreChannel tape drives, providing logical mapping of tape drives to multiple hosts. This provides the ability to share tape drives among multiple hosts and the ability to easily add or delete drives without reconfiguring the hosts.

NERSC/Probe tested the striping of FibreChannel tape drives. The SN6000 and attached drives were a convenient platform for these tests. NERSC saw approximately linear speedup going from one drive to three drives.

Because StorageTek decided not to commercialize RAIT technology, ORNL returned their SN6000.

#### **5.16.7 Storage Area Networks (SAN)/IP**

Genroco beta-tested a device that bridges FibreChannel disks to Gigabit Ethernet interface cards and uses the Scheduled Transfer (ST) protocol. It was tested at ORNL using Sun T300 disks and the Compaq DS20 server. In the testing, the server performed I/O to the FibreChannel disks through the server’s Gigabit Ethernet interface (rather than through its FibreChannel interface) exhibiting transfer rates up to 54 megabytes/second. As products of this type mature, high-performance FibreChannel equipment will be available through less-expensive and more easily-managed Ethernet switches and interface cards. ORNL is also testing such transfers using another Genroco device, a Gigabyte System Network bridge with Gigabit Ethernet and FibreChannel blades. See <http://www.csm.ornl.gov/PROBE/iSR.html> for details.

#### **5.16.8 Test of Gigabit Ethernet/FibreChannel Bridge**

The work to investigate bridging FibreChannel to Gigabit Ethernet (see 5.16.7) was a preliminary to studying SANs using IP. Early products have entered the market; some are tuned for local area networks and others for wide-area networks. They hold promise for inexpensive SANs and for wide-area transfers which do not require high bandwidth. NERSC did preliminary tests on a loaner iSCSI box from Cisco, testing transfers on a dedicated private network. They also did some wide-area-network tests.

### 5.16.9 Linux Hosts

NERSC began to evaluate the potential use of the Linux operating system and commodity Intel type PC hardware in a storage system environment. The evaluation verified support and stability of Linux with a multiprocessor Athlon system with multiport Gigabit Ethernet networking and FibreChannel and SCSI disk I/O subsystems. NERSC also compared performance one and two gigabit FibreChannel cards from Qlogic and LSI using with RAID devices from Sun Microsystems and RAID, Inc.

ORNL Linux work is centering on the use of Linux movers with HPSS in a variety of applications, including movers on PVFS nodes, movers on nodes using Web100/Net100 network tuning and movers on Cray network nodes.

### 5.16.10 New Storage Devices

In FY2002 new NAS devices and two gigabit FibreChannel disk arrays became available. NERSC tested both technologies and looked at new low-cost fiber attached disk arrays. Some results from those tests follow.

#### 5.16.10.1 FibreChannel Disk Array

NERSC ran timing tests on a StorageTek 9176 FibreChannel 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.

There is a paper on the web: [http://hpcf.nersc.gov/storage/hpss/probe/timing\\_stk/timing\\_stk.html](http://hpcf.nersc.gov/storage/hpss/probe/timing_stk/timing_stk.html) "Analysis of StorageTek 9176 FibreChannel Disk Array."

#### 5.16.10.2 EMC Clariion

NERSC tested the new EMC 2-Gigabit FiberChannel disk array. The chart below (Figure 5) shows results from testing large I/O transfers; NERSC also tested random small reads and writes.

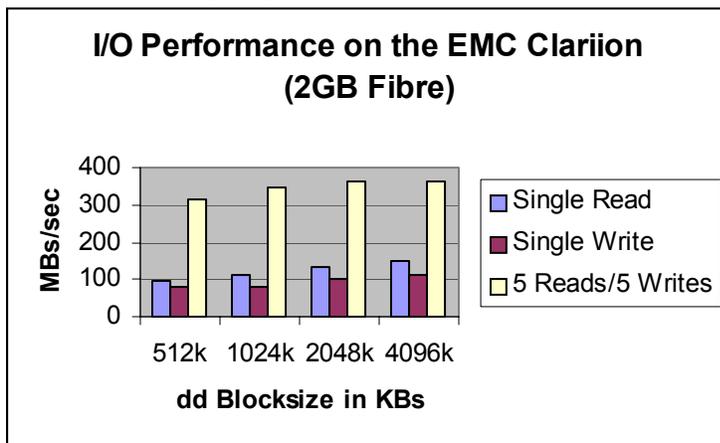


Figure 4. I/O Performance on EMC Clariion.

### 5.16.11 Low Cost Disk Arrays

Both sites evaluated a 1-terabyte FibreChannel RAID disk array priced under \$7000, the Condor system from RAID, Inc. Early ORNL tests demonstrated good performance and resilience to heavy use, leading ORNL to subsequently purchase seven more units for use in SciDAC projects. NERSC subsequently tested performance, with the results shown in Figure 6.

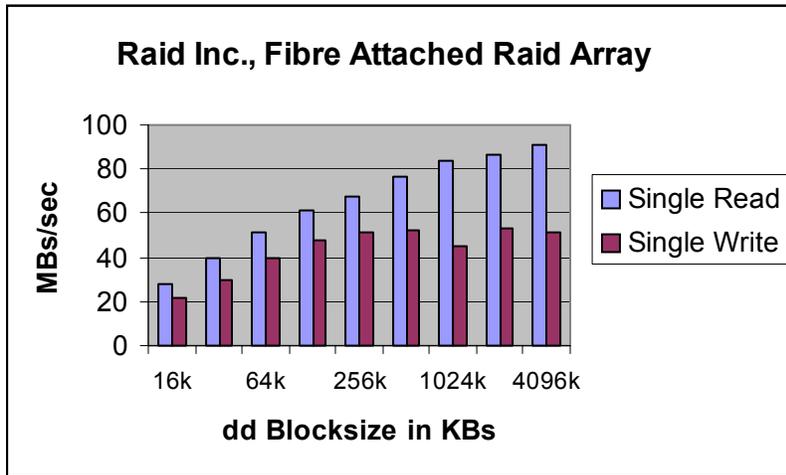


Figure 5. Condor Performance.

### 5.16.12 iSCSI Tests

NERSC did preliminary testing on a Cisco SN 5420 iSCSI router. The SN 5420 Storage Router provides servers with IP access to storage through SCSI routing using the iSCSI protocol. Some of the results were:

- Approximately 40% less performance using the iSCSI switch compared to a direct fibre connection between the host and a Sun T3.
- Poor performance figures in the WAN configuration, demonstrating the need for further tuning for this device.

### 5.16.13 Two-Stripe T300-Gigabit Ethernet Jumbo-Frame Performance

ORNL has two Sun/MaxStrat T300 FibreChannel RAID arrays, two FibreChannel interfaces in the RS/6000 Model S80, and two Gigabit Ethernet interfaces in the S80 and in the RS/6000 Model H70. Thus there is sufficient hardware to study the throughput of a two-stripe HPSS file transfer between the two nodes. Using the Hierarchical Storage Interface (HSI) application, two stripes could be read at 115 megabytes/second and written at 88 megabytes/second. Corresponding rates with parallel FTP were 92 and 58 megabytes/second, respectively. See <http://www.csm.ornl.gov/PROBE/2stripe.html> for details.

#### **5.16.14 StorageTek SCSI 9840 Tape Drive Testing**

ORNL also tested StorageTek SCSI 9840 tape drives. The drives were used with HPSS 4.1.1.1 over AIX 4.3.3 on the RS/6000 Model S80. Prior to these tests the drives had not been used on an S80 or with HPSS running over AIX 4.3.3. The goal of the test was to verify correct operation, and that was observed. These drives were added to the ORNL production system. See <http://www.csm.ornl.gov/PROBE/9840.html>.

#### **5.16.15 Configuration of StorageTek 9840 FibreChannel Tape Drives.**

There are several elements involved in supporting FibreChannel tape drives involves a number of elements, including hardware interfaces, the operating system, software "drivers" for FibreChannel, and software drivers for tape drives. There are multiple sources for the interfaces and the drivers. After extensive research and many trials, complicated by a dearth of documentation, ORNL successfully configured StorageTek 9840 drives and made them operational. An HPSS Operational Service Bulletin documenting the configuration process was provided at IBM/Houston's request. The drives were added to the ORNL production system. See <http://www.csm.ornl.gov/PROBE/Pprojects.html> for more information.

#### **5.16.16 Establishing an HPSS System on RS/6000 S80**

At the time of procurement, the RS/6000 Model S80 computer required the AIX 4.3.3 version of the operating system; HPSS had not been tested on that AIX release. At ORNL, HPSS was successfully compiled, installed, configured, and tested on the S80. The S80 was subsequently used to compile the latest releases of HPSS (HPSS 4.1.1.4 and then 4.2), to test FibreChannel disks and tapes, and to test HPSS movers linked with a Gigabit Ethernet jumbo frame. Throughout its life, ORNL upgraded the S80's HPSS system immediately after each patch or major release. See also <http://www.csm.ornl.gov/PROBE/aix433.html>.

#### **5.16.17 Establishing an HPSS system on AIX 4.3.3 on an IBM H70**

At NERSC, a first installation of HPSS on AIX 4.3.3 on an IBM H70 was performed, including compilation, installation, configuration, and testing. This installation was subsequently used to evaluate FibreChannel disk arrays and the IBM LTO library and tape system.

#### **5.16.18 StorageTek FibreChannel Equipment Installation**

In preparation for testing of HPSS movers, ORNL purchased FibreChannel disk equipment (including disks, storage processors, FibreChannel switch, and host interfaces for IBM, Compaq, Sun, and SGI computers) from StorageTek. The installation and configuration processes were so challenging that the work became a project in itself. A description of the process, notes taken during the work and the final documentation have been posted on the Web. See <http://www.csm.ornl.gov/PROBE/fiber.html> for lessons learned and configuration summaries.

## **6. SC2001 ACTIVITIES**

### **6.1 INTER-HPSS FILE TRANSFER DEMONSTRATION**

The ORNL booth included a demonstration of HSI in its HPSS-HPSS mode. Temporary accounts were established at ORNL, NERSC, San Diego Supercomputer Center 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. For instance, a transfer of file ABC from ORNL to SDSC was initiated by the HSI command “cp O:ABC S:”, where O: and S: represent ORNL and SDSC respectively.

### **6.2 HPSS WIDE-AREA REMOTE-MOVER DEMONSTRATION**

NERSC demonstrated 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. One version used the standard mover-mover protocol. The other was an experimental version that eliminated some handshaking. The end nodes used the Linux operating system. The GUI Web100 interface was used tune window sizes in real time 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 will then be 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. PRODUCTION-RELATED PROJECTS

### 7.1 PROJECTS WHICH HAVE RESULTED IN PRODUCTION IMPLEMENTATION

A large number of the research projects described in the previous sections have been put into production at one or more sites. We include this section to highlight the production value of Probe work.

- HSI.  
Probe-sponsored improvements in the base product are in use at several HPSS installations. Section 5.2.
- HPSS-HPSS transfers  
Such transfers are now available at any site running current HSI software. The capability was demonstrated at SC2001. Sections 5.2, 6.1
- Jumbo-frame Gigabit Ethernet.  
Probe testing verified that the ORNL supercomputers and HPSS could communicate correctly and quickly using jumbo frames. HPSS and the supercomputers at ORNL now communicate with one another using jumbo frames. Section 4.1
- FibreChannel disks and Gigabit Ethernet in HPSS.  
Prior to ORNL/Probe testing, no site had verified correct HPSS operation using this equipment. A number of sites, including ORNL, now use both technologies with HPSS in production. Section 5.5.
- Improve ORNL-NERSC bandwidth  
Protocol changes and improved connectivity combined to improve production transfer rates between the sites by a factor of 50. Section 5.1.
- Testing of new HPSS releases  
Some devices are not available in the IBM/Houston testbed. Testing at ORNL and NERSC on such equipment has qualified HPSS for use with those devices, which are in wide production use. Sections 5.5, 5.6, 5.8, 5.16.4, 5.16.14.
- Metadata performance testing  
Prior to deciding to replace the HPSS metadata engine, testing at ORNL verified that alternative approaches were much faster than the existing engine. The resulting release of HPSS (due in 2003) incorporates DB2 as the new metadata database. Section 5.3.

O SCSI-FibreChannel Bridge

One or two years ago, most high-performance tape drives used SCSI interfaces. Cabling issues and backplane space encourage FibreChannel deployment. Probe tested, and ORNL put into production, a SCSI-FibreChannel bridge to connect eight IBM tape drives to two fibrechannel interfaces. Section 5.16.2.

- LTO

Work done at NERSC to develop and integrate LTO equipment into HPSS has been a part of the HPSS product since Release 4.3. Section 5.16.4.

- Remote Movers

This project showed the benefit of having movers in each site's HPSS configuration that were physically located at the other site. When both installations are at the same release of HPSS we plan to implement the capability. Section 5.7.

## 7.2 PROJECTS WHICH DISCOURAGED PRODUCTION IMPLEMENTATION

Some technologies look attractive on paper but turn out to be inadequate. Results from testing in Probe showed that the following technologies were not sufficiently valuable for production. Probe research may have avoided unnecessary effort or expenditure.

- GSN transfer rates are very attractive and ORNL/Probe testing set two performance records. However, no computers with PCI backplanes are capable of driving transfers at full GSN rates. That fact, combined with the per-port expense of GSN equipment, caused ORNL to abandon GSN. Section 5.16.5.
- The Scheduled Transfer protocol avoids some copying of data within node memory and consequently speeds transfers. However, it proved to be very complex and to have only rudimentary error handling capabilities. It was entirely unsuited for wide-area use so the thrust was abandoned. Section 5.14.
- SAN/IP is a popular current topic. ORNL and NERSC have done some work with the technology, but it appears inadequate for our needs. Its target market is small or low-budget installations or moderate-throughput applications, neither of which characterize us. Section 5.16.12.
- Texas Memory Systems markets a memory device that emulates disks. Testing on the device went very well, but the device did not provide benefits to HPSS. While I/O operations were fast, the device is so expensive that it is appropriate (within HPSS) only for metadata processing. HPSS metadata processing is not bottlenecked by I/O; the TMS equipment did not help enough to justify its cost. No such devices (TMS or competitors) have been sold to HPSS installations. Section 5.16.1.

## 8. PUBLICATIONS AND PRESENTATIONS

### 8.1 PUBLICATIONS

#### 8.1.1 Refereed Journal Papers:

- 1) N. F. Samatova, A. Geist, G. Ostrouchov and A. Melechko (2003.c), "Parallel Out-of-core Enumeration of Metabolic Pathways," *Special Issue on High-Performance Computational Biology, Journal of Parallel and Distributed Computing: An International Journal* (invited paper, to appear).
- 2) 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).
- 3) N. F. Samatova, G. Ostrouchov, A. Geist, A. Melechko (2002.a), "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.
- 4) 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.

#### 8.1.2 Refereed 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 International Parallel and Distributed Processing Symposium (IPDPS.02)* pp.8.
- 2) N. F. Samatova, G. Ostrouchov, A. Geist, A. V. Melechko (2001.a), "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) Y. Qu, G. Ostrouchov, N. F. Samatova, A. Geist (2002), "Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets," *Workshop on High Performance Data Mining at The Second SIAM International Conference on Data Mining*, p.4-9.
- 4) F. N. Abu-Khzam, N. F. Samatova, G. Ostrouchov, M. A. Langston, and G. A. Geist (2002), "Distributed Dimension Reduction Algorithms for Widely Dispersed Data," *Fourteenth IASTED International Conference on Parallel and Distributed Computing and Systems*.
- 5) 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.
- 6) 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.

#### 8.1.3 Other Publications

- 1) 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, "Probe Project Status and Accomplishments – Year Two," ORNL/TM-2002/62, February 1, 2002.
- 2) 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.

- 3) Harvard Holmes, "Exploration of Cache Behavior Using HPSS Per-File Transfer Logs," Lawrence Berkeley, LBNL-49330, November 2001, <http://hpcf.nersec.gov/storage/hpss/probe/caching/cache-behavior.pdf>.
- 4) R. D. Burris, M. K. Gleicher, H. H. Holmes, N. L. Meyer, D. L. Million, "Probe Project Status and Accomplishments," ORNL/TM-2001/25, February 1, 2002.

#### **8.1.4 Submitted/Written Papers**

- 1) N. F. Samatova, F. N. Abu-Khazan, D. Bauer, G. Ostrouchov, M. A. Langston, and A. Geist (2003.a), Heuristic Minimum Spanning Tree Algorithms for Distributed Data Sets.
- 2) N.F. Samatova, F. N. Abu-Khazan, G. Ostrouchov, M. A. Langston, and G. A. Geist (2003.b), Distributed and Robust Linear Time PCA Algorithm.
- 3) G. Ostrouchov, N. F. Samatova, Y. Qu, A. Geist (2003.a), Principal Component Analysis for Distributed Datasets.
- 4) G. Ostrouchov, J. Hespden, N.F. Samatova (2003.b), PCA-based Data Reduction Algorithm.
- 5) N. F. Samatova, G-X Yu, P. Chandramohan, H. Park, G. Ostrouchov, A. Geist, N. Maltsev (2003.d), Hierarchical feature extraction based approach to functional differentiation of highly homologous protein functional groups.

#### **8.2 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".

#### **8.3 Presentations, Posters and Demonstrations**

- 1) S. Cholia, "Web Based Reliable File Transfers Between HPSS-Grid FTP Systems," demonstration at SC2002, Baltimore, Maryland, November, 2002.
- 2) N. L. Meyer, " HPSS @ NERSC 6/2002," presentation to the HPSS Users Forum, Indiana University, Indianapolis, Indiana, June 2002.
- 3) R. D. Burris, "Oak Ridge National Laboratory HPSS Site Report – 2002," HPSS Users Forum, Indiana University, Indianapolis, Indiana, June 2002.
- 4) N.F. Samatova, R. Burris, G. Ostrouchov, and T. Potok (2002), Scientific Data Management. *ASCR/ASCI CS Review, April 1, 2002.*
- 5) N.F. Samatova, A. Geist, G. Ostrouchov, and A. Melechko, "Parallel Out-of-core Algorithm for Genome-Scale Enumeration of Metabolic Systemic Pathways," First International Workshop on High Performance Computational Biology at the International Parallel and Distributed Processing Symposium (IPDPS.02), Monday, April 15, 2002, Fort Lauderdale, Florida.
- 6) H. H. Holmes, "Probe: A Research and Development Environment to Meet the Performance Challenge," presentation to the Directors Review, LBNL, Berkeley, California, April, 2002
- 7) N. F. Samatova, G. A. Geist, and G. Ostrouchov, "RACHET: Petascale Distributed Data Analysis Suite," *SPEEDUP Workshop on Distributed Supercomputing Data Intensive Computing*, March 4-6, 2002, Leukerbad, Valais, Switzerland.
- 8) N. F. Samatova, "Advanced Algorithms for Computational Biology," *Science and Technology Review*, February 11, 2002, ORNL.
- 9) N. F. Samatova, G. Ostrouchov, A. Geist, "Scientific Data Mining Research at CSM," presented to Life Sciences Division, ORNL, January 11, 2002.

- 10) T. H. Dunigan, "Net100 measurement and tuning," Internet2 workshop, Tempe, January 2002.
- 11) T. H. Dunigan, "Net100," DOE SciDAC PI meeting, Wash. DC, January 2002.
- 12) R. D. Burris, "Probe Data Storage, Transfer and Research Facility," presentation to networking workshop sponsored by Thomas Ndousse, December 2001.
- 13) T. H. Dunigan, "Net100," DOE network workshop, Oak Ridge, November 2001.
- 14) T. H. Dunigan, "Net100 project," Network BOF, SC 2001, Denver, November 2001.
- 15) 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.
- 16) 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 IUVESTA 15th International Vacuum Congress (IVC-15), AVS 48th International Symposium (AVS-48), 11th International Conference on Solid Surfaces (ICSS-11), San Francisco, October, 2001.
- 17) N. F. Samatova and G. Ostrouchov, "Scientific Data Mining Research under Probe," presented to Distributed Computing Group at CSMD/ORNL, September 18, 2001.
- 18) N. F. Samatova and G. Ostrouchov, "Multi-agent based High-Dimensional Cluster Analysis," presented to Steve Eckstrand, OS/DOE, August 9, 2001.
- 19) B. Palsson, "Predictive models of biochemical pathways and microbial behavior," GTL/DOE Workshop, August 7, 2001 (minor contribution).
- 20) 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.
- 21) 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.
- 22) T. H. Dunigan, "Net100 overview," Web100 conference, Boulder, CO, July 2001.
- 23) T. H. Dunigan, "Web100 testing at ORNL," Web100 conference, Boulder, CO, July 2001.
- 24) R. D. Burris, D. L. Million, "Probe Plans and Status," presentation to the SciDAC SDM ISIC Kickoff, July 2001.
- 25) M. K. Gleicher, "Texas Memory Systems Testing," presented to the HPSS Users Forum, June 2001.
- 26) M. K. Gleicher, "HSI," presentation to the HPSS Users Forum, June 2001.
- 27) R. D. Burris, "ORNL/Probe Performance Tests," presentation to the HPSS Users Forum, June 2001.
- 28) N. L. Meyer, "HPSS @ NERSC 6/2001," presentation to the HPSS Users Forum, San Diego Supercomputer Center, La Jolla, California, June 2001.
- 29) Wayne Hurlbert, S. Cholia, N. Johnston and M. Andrew, "Performance @ NERSC 6/2001," presentation to the HPSS Users Forum, San Diego Supercomputer Center, La Jolla, California, June 2001.
- 30) 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.
- 31) N.L. Meyer, "PROBE", presentation to Dan Hitchcock, LBNL, Berkeley, California, January, 2001.
- 32) N. F. Samatova, "Vector Space Model for Lean Cell Formation," presented at CSMD seminar, December 1, 2000.
- 33) SC2001 Poster, "Rachet: Petascale Distributed Data Analysis Suite."

- 34) SC2001 Poster, "Cracking Computational Complexity for Genome Scale Modeling of Metabolic Pathways."
- 35) SC2001 Demonstration, "Cracking Computational Complexity for Genome Scale Modeling of Metabolic Pathways."
- 36) ORNL is referenced as collaborating in the improvement of HSI in the on-line publication NPACI and SDSC Online dated September 20, 2000.
- 37) R.D. Burris, presentation to the ESnet Steering Committee, September 13, 2000.
- 38) R. D. Burris, D. L. Million, S. R. White, M. K. Gleicher, and H. H. Holmes, Poster Paper at the High Performance Distributed Computing conference in Pittsburgh, PA, August 1, 2000.
- 39) M.K. Gleicher, Special Presentation: HIS to the HPSS User's Forum, July 25, 2000.
- 40) R.D. Burris, Presentation to Presentation to the HPSS User's Forum, July 26, 2000.
- 41) R. D. Burris, Presentation to Fred Johnson in Oak Ridge, June 20, 2000.
- 42) R.D. Burris and H.H. Holmes, Presentation to the ESCC Meeting, April 25-27, 2000, presented by H. H. Holmes.
- 43) R.D. Burris, Presentation to Dan Hitchcock in Oak Ridge in April 2000.
- 44) R.D. Burris, Presentation to the Eighth Goddard Conference on Mass Storage Systems and Technologies, March 29, 2000.
- 45) R.D. Burris, Presentation to the HPSS User's Forum, September 1999.
- 46) R.D. Burris, Presentation to the HEPiX in October 1999.

## 9. SUMMARY

Probe evolved from a simple testbed, early in its life, to a productive research facility with notable accomplishments in 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. Its facilities continue to be used as a key component of a variety of MICS-funded activities, including both base-funded projects and SciDAC projects such as the SDM ISIC, the Terascale Supernova Initiative and the DOE Science Grid and Earth Systems Grid II.

The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

**APPENDIX A  
ORNL EQUIPMENT DESCRIPTIONS**

<b>Node Name</b>	<b>Machine</b>	<b>OS</b>	<b>Processor s</b>	<b>Memory MB</b>	<b>Disk GB</b>	<b>Network</b>	<b>Software</b>
<b>marlin</b>	IBM H70	AIX	4	2048	280+	GigE	C, Fortran, DB2, Oracle
<b>earl</b>	IBM B80	AIX	2	1024	36	FastE	C
<b>satchel</b>	IBM 44P-170	AIX	1	1024	54	FastE	C
<b>sneezy</b>	IBM 44P-170	AIX	1	512	310	FastE	C
<b>bucky</b>	IBM 44P-170	AIX	1	1024	54+	FastE	C
<b>bashful</b>	IBM 44P-170	AIX	1	512	27	FastE	C
<b>jupiter</b>	IBM p630	AIX	4	4096	146+	GigE	C
<b>saturn</b>	IBM p630	AIX	2	2048	146	GigE	C
<b>neptune</b>	IBM p630	AIX	2	2048	146	GigE	C
<b>happy</b>	Sun E450	Solaris	1	512	108	FastE	C/C++, OPNET
<b>sleepy</b>	Sun E250	Solaris	2	512	430+	GigE	C/C++, HRM
<b>dilbert</b>	Intel	Linux	2	512	240	FastE	C
<b>wally</b>	Intel	Linux	2	512	240	FastE	C
<b>alice</b>	Intel	Linux	2	512	240	FastE	C
<b>phb</b>	Intel	Linux	2	512	240	FastE	C
<b>Alice cluster</b>						GigE	The four previous nodes clustered using OSCAR
<b>Farkle cluster</b>	AMD	Linux	1/node	512	480/ node	GigE	Copy of Argonne cluster for PVFS
<b>bubba</b>	Intel Xeon	Linux	4	8192	260	GigE	C
<b>laurabeth</b>	Intel Xeon	Linux	2	4096	260+	GigE	C
<b>lindasue</b>	Intel Xeon	Linux	2	4096	260	GigE	C
<b>maryjo</b>	Intel Xeon	Linux	2	4096	260	GigE	C
<b>sallyjean</b>	Intel Xeon	Linux	2	4096	260+	GigE	C

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



**APPENDIX B**  
**NERSC EQUIPMENT DESCRIPTIONS**

---

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

---

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