Scalable File Systems for High Performance Computing

Darrell D. E. Long, Scott A. Brandt, Ethan L. Miller
Patrick E. Mantey, Alexandre Brandwajn and Katia Obraczka
Center for Storage Systems Research
University of California, Santa Cruz

June 1, 2001

1 Executive Summary

Simulations on high performance computer systems produce very large data sets. Rapid storage and retrieval of these data sets present major challenges for high-performance computing and visualization systems. Although computing speed and disk capacity have both increased at exponential rates over the past decade, disk bandwidth has lagged far behind. Moreover, existing file systems for high-performance computers are generally poorly suited for use with workstations, necessitating the copying of data for use with visualization systems. Our proposed project will address research issues in the design of a high-performance multi-petabyte storage system directly targeted for use in post-Purple computing systems planned for Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), and Sandia National Laboratory (SNL).

The goal of our research is a storage system architecture that supports both high-speed capture of data from massive simulations and the interactive use of this data by investigators via visualizations. Our research will be driven by the data rates and usage scenarios that characterize the mission of the Defense Programs (DP) Laboratories. Results of this research will be validated and demonstrated by creation of prototype implementations of this architecture. A major value of the results of this research will be its use by the DP Laboratories in providing direction to the vendors of post-Purple systems in building new file and storage systems.

We propose to improve both file system performance and functionality by building a storage system from object-based storage devices (OBSDs) connected by high-speed networks. A key advantage of OBSDs in a high-performance environment is the ability to delegate low-level block allocation and synchronization for a given segment of data to the device on which it is stored, leaving the file system to decide only on which OBSD a given segment should be placed. Since this decision is quite simple and allows massive parallelism, each OBSD need only manage concurrency locally, allowing a file system built from thousands of OBSDs to achieve massively parallel data transfers. Additionally, OBSDs can each manage their own storage consistency, removing the need to run a system-wide consistency check that could take days on a petabyte-scale traditional file system.

Our OBSD-based file system will be based upon a flat name space, employing a cluster of metadata servers to translate "human-readable" names into a file identifier in this flat namespace. The use of multiple servers rather than one provides redundancy and higher performance, allowing the metadata service to scale with the number of file system clients. These servers can provide multiple translation mechanisms, including database lookup as well as more conventional hierarchical directory structures. Security is also an important design consideration for file systems; we will explore security schemes that may include end-to-end encryption to protect data as directed by standard access control mechanisms, and provide unified access control mechanisms among the several systems that share the file system.

Interconnecting all of these OBSDs is another issue that we will explore. A single OBSD will likely transfer data at less than 1 GB/s, thus requiring hundreds or thousands of OBSDs to satisfy the bandwidth requirements of large-scale computations. Routing data to the correct OBSD and ensuring that there is sufficient network bandwidth for the data to be transferred will be crucial to ensuring that the file system works at the anticipated data rates.

Because a petabyte storage system will be shared between high-performance computing engines and researchers doing visualization and program development, our research will explore issues that arise from such sharing. The storage system must handle simultaneous requests from multi-teraflop (and eventually petaflop) computers, visualization systems, and individual workstations. These requests will differ greatly in terms of file size distribution and access pattern, but the storage system must handle them all efficiently. For example, scheduling may be necessary to insure that visualization applications can proceed uninterrupted while data is being stored by a long-running computation. Moreover, these clients may not be collocated, so the storage system must deal with issues of wide area access and caching.

Addressing this problem is of crucial importance to the success of the stockpile stewardship role of the National Laboratories because storage system performance will be the bottleneck to faster computation. Improving file system performance will directly result in faster computations and more usable visualization systems. The ability to use commodity components will reduce overall storage system cost by leveraging the huge market in commodity storage and computational platforms.

While conducting this research, we will work closely with researchers at the three national laboratories to help focus our research directly on problems facing ASCI. For example, our file system will have a POSIX-compliant interface to allow standard code and libraries on the diverse computing platforms used in these laboratories to directly access the shared data. We will work with the researchers at the Defense Programs (DP) National Laboratories to understand the varied I/O models that are currently in use, and how these will affect the design of our file system. We will discuss issues such as security with people at at all three DP Laboratories, taking particular advantage of our close proximity to LLNL, allowing us to propose solutions that are both efficient and usable in their environments.

We are currently working with researchers at LLNL on several storage-related projects, including an ASCI Level 3 project to explore scalable file systems and an ISCR project to examine the use of probe-based storage in high-performance file systems. Should this proposal be funded, we will fold our level 3 project on object-based storage devices into the level 2 collaboration. We will continue to work with LLNL and the other DP Laboratories on our proposed research. Because of our proximity to Livermore, we will be able to attend meetings, present seminars, and exchange students as desired.

To accomplish the research described herein, we envision seven primary tasks: Tri-Lab Interaction/Integration Plan, Workload Characterization, Metadata Server Software, OBSD Software, Client File System Software, Interconnection Network, and Analysis. These tasks are described fully in Section 2.4.

At the University of California, Santa Cruz we have brought together a unique collection of talent in the Storage Systems Research Center (SSRC). Our center is unique due to the number of academic researchers involved in storage research and its systems-oriented focus. Specifically, our research covers high-performance storage architectures, structures, algorithms, security, and software. The SSRC includes a number of experienced researchers with research interests covering many aspects of storage systems relevant to the proposed work. Prof. Darrell Long, the Director of the SSRC, has 15 years experience in storage systems and is one of the architects of the IBM Storage Tank system, with more than 80 publications in storage and other areas of computer systems. Prof. Patrick Mantey, former Dean of the Jack Baskin School of Engineering and current Director of the Organized Research Unit (ORU) that includes the SSRC, has broad experience with computer systems both at IBM and in academia. Prof. Scott Brandt has 17 years of academic and industrial research and development experience in computer systems, including 25 publications and six patents. Prof. Ethan Miller has 10 years experience and 40 publications in various aspects of storage systems. Profs. Alexander Brandwajn and Katia Obraczka round out the team with significant expertise in systems analysis, storage system performance, and high-speed networking. Prof. Brandwajn has over 20 years of experience spanning industry and academia, and in systems analysis was General Chair of the ACM SIGMETRICS 2000 Conference.

2 Technical Discussion

The ASCI program provides unique challenges for researchers that are not found in any other environment. The post-Purple requirement of 2 PB (petabytes) of data delivered at a sustained 100 GB/s–1 TB/s is beyond what is even contemplated today by most researchers. Scalability is the key to achieving this goal. While storage components will continue to increase in density and in speed, they will not improve quickly enough to meet this goal without a high degree of parallelism. We believe this goal can best be achieved sing a highly scalable architecture made up of commodity components.

2.1 Introduction

Storing and retrieving large quantities of data at high speed is a major challenge for high performance computing systems. Though both computing speed and disk capacity have increased at exponential rates over the past decade, disk bandwidth has lagged far behind. Moreover, existing file systems for high-performance computers [11, 19] are generally poorly suited for use with workstations, necessitating the copying of data for use with visualization systems.

Next generation high-performance multi-petabyte storage systems targeted for terascale post-Purple computing systems will be shared between high-performance computing engines and researchers doing visualization and program development. They will have very high bandwidth local access as well as lower bandwidth wide-area access. Accesses will be done simultaneously by hundreds or thousands of concurrently executing nodes, and may be to a single very large file or many smaller files with throughputs of over 100 GB/s. Finally, although always important, wide-area access will make access control an even greater concern than in precursor systems.

Local access is primarily characterized by two access patterns generated by large-scale simulations used for the stockpile management mission of the ASCI program. In the first access pattern, nodes of a scalable multi-computer issue high throughput writes to a small number of very large files. Each multi-computer node following the second access pattern writes large quantities of data to its own files; in this case, nodes do not write to shared files. The files written by large scalable systems are read by visualization and analysis systems and should also be available for wide-area access.

Current-generation systems have throughput requirements on the order of 1–10 GB/s from hundreds of concurrently executing processor nodes executing large-scale simulations. The simulations are highly write-dominated and I/O performance and storage system write performance is critical. File creation times may also be important in those simulations where each node writes to a separate file. In contrast to simulations, visualization applications are highly read-dominated and may generate highly variable access patterns. Overall throughput requirements are similar to those for simulation applications, but read performance is critical for these applications.

High-performance metadata management is also very important. Simultaneous writes to a single file must proceed in parallel without excessive lock or other metadata management overhead. Similarly, bursts of file create operations must not create a bottleneck in the metadata management. Highly random read accesses generated by simulation engines must proceed at similar rates.

While the access patterns will be more or less the same, next generation terascale computing systems will have throughput and other performance requirements 10–100 times greater than that of current systems. Where current systems have throughputs on the order of 1 GB/s overall, next generation systems will have throughput requirements of over 100 GB/s overall.

Wide-area access is currently in the form of whole file transfers via indirect methods such as FTP. Next generation systems should support direct remote file access, although the bandwidth requirements will be far lower than for the local access by simulation and visualization engines. Support for wide-area access will require the ability to support standard file system interfaces as general wide-area systems cannot be expected to run special-purpose file systems solely to access this data.

Furthermore, the need for wide-area access increases the need for robust security. Security and access control issues are of primary importance in post-Purple National Laboratories storage systems. Local access to files will have to be controlled based on existing security standards. However, with direct wide-area access, security becomes even more important.

2.2 Detailed Description

Just as terascale computers are built from relatively inexpensive components connected by high-speed networks, we propose to improve both file system performance and functionality by building a storage system from object-based storage devices (OBSDs). As with the building blocks in a terascale computer, the OBSDs in our file system must be interconnected by high-speed networks; however, latency is less of an issue for the storage system because of the size of the data elements and the latency involved in reading data from disk.

The main hardware component of our file system design is an object-based storage device—one or more disks (or other storage devices) managed by a single CPU and seen by the file system as a single device, in a manner similar to network-attached secure disks (NASD) [8]. The data in the file system is distributed across many OBSDs, with high bandwidth coming from large numbers of concurrently operating OBSDs. Each OBSD is responsible for managing and allocating its own storage;

requests to an OBSD are of the form "write (or read) this range of bytes from file X," with low-level placement of the data and free space management left to the OBSD. High-level information such as the striping pattern across OBSDs and translation of names to file identifiers are left to a metadata server, which is accessed by the user only when a file is opened or closed. This proposed file system design is shown in Figure 1.

The key advantage of OBSDs in a high-performance environment is the ability to delegate low-level block allocation and synchronization for a given segment of data to the device on which it is stored, leaving the file system to decide only on which OBSD a particular segment should be placed. This decision is quite simple and allows massive parallelism; thus, each OBSD, shown in Figure 2, can manage concurrency locally. This allows a file system built from thousands of OBSDs to achieve massively parallel data transfers. Because each OBSD is managing its own storage allocation, it can also check its own consistency independently of other OBSDs in the file system, removing the need to run a system-wide consistency check that could take days on a petabyte-scale traditional file system.

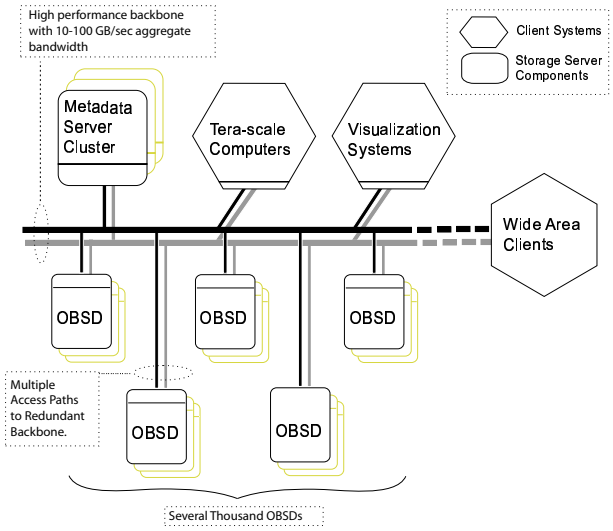


Figure 1: Proposed architecture.

Our OBSD-based file system will provide a flat name space, requiring the use of a metadata server to translate "human-readable" names into a file identifier. This server can provide multiple translation mechanisms, including database lookup as well as more conventional hierarchical directory structures. Security is also an important design consideration for file systems; we will explore security schemes that may include end-to-end encryption to protect data as directed by standard access control mechanisms, and provide unified access control mechanisms among the several systems that share the file system.

The file system research we will perform will be fully POSIX compatible, providing block-level interfaces to file system clients. Rather than including MPI-IO directly in the file system, we will have a user-level implementation of MPI-IO. This approach will maintain flexibility for systems that do not use MPI-IO while still allowing parallel programs to take advantage of the improved layout and data specification provided by MPI.

Because the petabyte storage system will be shared between high-performance computing engines and researchers doing visualization and program development, it will must handle several types of workloads simultaneously. Terascale computers will demand high-bandwidth access to large files, while program developers using the same storage system may read and write small files at a high rate. Moreover, it is likely that some of the researchers will not be collocated with the terascale computation engine on which the simulation is run, making it necessary for the file system to handle wide-area access efficiently. We will explore issues in wide area access and caching, as well as studying problems that may occur when high bandwidth and low bandwidth clients read and write files simultaneously. For example, scheduling may be necessary to insure that an interactive visualization application can be run without interruption while a long-running computation on a post-Purple computer stores its results. system does not interfere with a long-running computation on a post-Purple computer.

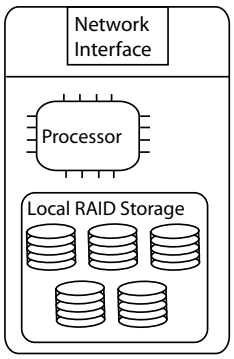


Figure 2: A single OBSD.

While conducting this research, we will work closely with researchers at the three DP National Laboratories to help focus our research directly on problems facing ASCI. For example, our file system will have a POSIX-compliant interface to allow standard code and libraries on the diverse computing platforms used in these laboratories to directly access the shared data. We will work with the researchers at the national laboratories to understand the varied I/O models that are currently in use, and how these will affect the design of our file system. We will discuss issues such as security with people at LLNL, SNL, and LANL, allowing us to propose solutions that are both efficient and usable in the labs' environment.

2.2.1 Metadata Server(s)

The metadata server is a key component of an OBSD-based file system because OBSDs store file data named by a unique file identifier and offset within a file. In order to provide a user-friendly POSIX interface, the file system must store metadata as well as data; storing it on a metadata server cluster separates data from metadata, allowing OBSD and metadata servers to be scaled independently [5]. This approach has other advantages beyond simple scaling, however.

The first advantage to a separate metadata server is the ability to use multiple, flexible name spaces. Certainly, a name

server will need to support a hierarchical directory structure similar to that used in Unix. This can be done in many ways, ranging from a standard Unix-like tree structure to an optimized database using B-trees. However, the flat name space on the OBSDs allows for multiple name spaces to coexist for the same data. For example, HPC users may want to organize their files using a more full-featured database, attaching many attributes to each file stored on the OBSDs. This database could then be queried for files matching particular criteria, allowing researchers to quickly find the data files they need. Because the system can coexist with other naming schemes, moreover, there is no requirement to convert programs to use database access; it can be provided as an adjunct naming scheme. Other schemes are also possible, such as a naming scheme that produces a list of candidates ranked by "fit" for the query provided to the metadata server.

Scalability of the metadata server is a major concern for several reasons. First, a multi-petabyte system will need to support millions of files ranging from a few bytes to a few terabytes. Additionally, the server must support tens of thousands of processes opening a single file simultaneously without bogging down excessively. It must also allow high-bandwidth access to files without the need to query the metadata server; otherwise, the server will become a bottleneck. In our system, clients need only access the metadata server on a file open or close; there is no need to contact it during file reads and writes, greatly improving scalability by drastically reducing the number of requests the metadata server must handle. This is particularly true for operations on large files. For example, a single 1 TB file that is written by 10,000 compute nodes would require that the metadata server service 20,000 requests—10,000 open requests and 10,000 close requests. If each node needed to contact the server just once per 8 MB chunk, however, the server would need to handle 128,000 additional requests; this number would increase for smaller chunks or larger files. Thus, this approach of separating naming metadata from positional and block allocation metadata greatly improves scalability by reducing the number of operations the metadata server cluster must handle. Additionally, separating metadata into its own server allows us to use a metadata server cluster that can be scaled to provide metadata services at as high a rate as may be required by the systems using it.

The metadata server is responsible for keeping information about files (size, ownership, security, etc.) and for maintaining high-level location information for each file. Rather than specify the location for each block, however, the metadata server provides clients with sufficient information to locate the OBSD that holds a particular block. This information is set by the file's creator, and allows file striping to be controlled on a per-file basis. If no striping is specified at creation, the system will supply a reasonable default.

Our research will also address caching, locking and consistency issues appropriate to the scale of the systems involved [4]. For example, caching of data may not be beneficial for multi-terabyte files accessed sequentially, but is very appropriate for regions of a file accessed by a thread in a parallel application. We will also address the issues introduced by having either a single multi-terabyte file shared by thousands of threads in a parallel application, or thousands of files each accessed by a different thread. We will investigate the best way to provide cache coherency and locking both to conform to the POSIX specification, and to provide the best performance for MPI-IO and the other I/O interfaces used at the DP laboratories.

2.2.2 Object-Based Storage Devices (OBSDs)

There are several major issues in the design of a single OBSD, including performance, dependability, and security. We will consider different on-disk layouts for data and location information on an OBSD, and evaluate their effectiveness at helping OBSDs achieve high performance and reliability.

One promising approach to organization of data on disk is the use of inverted indices. In this scheme, each block on disk is tracked by a single pointer indicating which file and offset apply to each block. This scheme can be tuned to keep blocks of a file contiguous on disk, increasing sequential read and write performance [16]. In addition, inverted indices can be implemented with relatively simple code, in contrast to more complex data structures such as B-trees. This simplicity will increase software reliability, making it unlikely that data will be lost or corrupted due to software errors. Another advantage of inverted indices is the ability to delay OBSD metadata writes indefinitely without risking the loss of file system consistency as can happen in many conventional file systems. During bursts of extremely high data transfer rates, the OBSD can safely delay writing metadata while streaming data to disk. Moreover, we can further reduce the window of vulnerability by placing metadata in temporary locations while streaming data. We will compare OBSD implementations using inverted indices to those using other file system indexing techniques, showing the benefits and shortcomings of each approach in building OBSDs.

Another issue plaguing existing file systems is file system dependability. Our research will focus on ways to ensure that data remains available despite hardware failure, and on ways to increase the reliability of file system software. Clearly, it is impossible to run a serial consistency check over an entire multi-petabyte file system. However, it is possible to check the data structures for an individual OBSD without outside help, allowing consistency checks to proceed in parallel. Since the metadata server only maps blocks to OBSDs statically, it does not contain any dynamically modified placement information, and can thus itself be checked without references to an OBSD. Because all of this checking happens in parallel, total time to check the file system is that required to check a single OBSD. Reducing the number of checks is important as well, though, and inverted indices can will greatly reduce the chance that such a check will be necessary in the first place. Another issue related to dependability is the ability of a file system to recover from accidental or deliberate file deletions. We will incorporate the

ability to take and keep snap-shots into our system, allowing recovery from deletions without the need to recover data from a backup. This is a crucial factor because it will be impossible to back up a file system with several petabytes of data written at 100 GB/s.

Dependability includes other factors beyond software, such as the intermittent or permanent loss of hardware. Each OBSD will use RAID, allowing it to survive a single disk failure. The failure of an entire OBSD, however, is a different matter: with thousands of OBSDs, we expect that OBSDs will experience transient failures that result in data being unavailable temporarily. The file system must continue to operate despite the loss of an OBSD, storing new data destined for the failed OBSD on a hot spare and merging the spare with the original when it comes back up. We will explore mechanisms for inter-OBSD redundancy, allowing the system to provide full access to data despite the loss of one or more OBSDs.

2.2.3 Interconnection Network

One of the main challenges of this research is to be able to achieve aggregate data rates of over 100 GB/sec. Our approach is to take advantage of the intrinsic massive parallelism of the proposed architecture. For example, if we partition the data across 500 OBSDs, 100 GB/sec aggregate bandwidth can be achieved by transferring data in parallel from/to each OBSDs at 200 MB/sec, or 1.6 Gbit/sec. This rate cannot be achieved with current Gigabit Ethernet technology. If we use 10 Gigabit Ethernet, currently being developed by a number of different vendors, our design can reach 100 GB/sec with just 500 OBSDs. Alternatively, we could use 2000–5000 or more OBSDs, reducing the bandwidth required per OBSD to well under 1 Gbit/sec. Combining these approaches will result in a storage system able to handle the expected demand of 100 GB/sec or more.

In order to meet ASCI storage requirements, our interconnection network will allow parallel and concurrent access to data across any number of OBSDs in the system. In the case of write-intensive large-scale simulations involving very large data items, the high bandwidth requirements are met by having data blocks striped across OBSDs and accessed in parallel. For simulations involving smaller files, the interconnection network must allow concurrent access to multiple files.

One possible approach is to use a point-to-point crossbar switching fabric to interconnect the OBSDs and the high-performance processor elements. An example of this type of high-speed interconnection network is the point-to-point crossbar inside Hitachi's HDS 9900 Storage Controller [20]. The nominal 0.8 GB/sec bandwidth per board of the HDS 9900 crossbar switch would be sufficient to interconnect the ODSDs to the processor elements.

As with the other hardware components of the proposed storage system, our intention is to build the interconnection network entirely out of COTS components. The HDS 9900 system just described demonstrates that, using current technology, it should be feasible to build an interconnection network whose aggregate bandwidth meets the bandwidth goals of the system. As part of this project we will investigate trends in current interconnection network technology in order to design and prototype the fabric interconnecting ODSBs and processing elements. The goal is not only to meet current bandwidth requirements but also to be able to scale to larger number of processors and OBSDs and even higher bandwidths.

2.2.4 Wide Area Access

Wide area access to the petabyte file system is as important as high-bandwidth local access because it is one of the primary methods of visualizing data stored on the system. There are two separate issues with wide area access: streaming and prefetching data to reduce latency, and caching in a wide area file system.

For large data sets, prefetching and streaming are essential to keep high-performance visualization systems running without jitter, but it is not clear that the available bandwidth will be sufficient to do this. Moreover, such a system should predict file usage and prefetch data to reduce both transmission latency and latency due to congestion. We are currently investigating mechanisms for predicting file usage for prefetching and other clustering [1], and will apply this research to prefetching large data sets, potentially spread among more than one file, across a wide area network.

Caching in a wide area network is a second major issue with wide area file systems, applying to a different set of files than streaming. Small files can be cached in a wide area file system, and such caching is essential to maintaining high performance. Reducing the demand for remote files improves response time by eliminating round trip travel and by reducing the load on the remote storage system. However, caching may result in inconsistencies in the file system. We have significant expertise in data replication, with members of our research having made significant contributions to the field [9, 18, 17]. Depending on the characteristics of the workload, it may be possible to apply locking techniques to this problem or it may require more complex consistency mechanisms. The most important challenge is to provide the performance necessary to satisfy the ASCI mission, while doing it at the least possible cost.

2.2.5 Security

Security is a another issue that our research will address. POSIX standards include requirements for access control, but do little to ensure that data remains safe on the disk or in transit. We will explore ways of doing end-to-end encryption of data, using

both hardware and software to ensure that data exists in the clear only on clients requesting the data. This approach will be transparent to file system users, both in terms of usage [15, 7] and performance [6], and will provide a high level of security while adhering to POSIX standards. In particular, though encryption at high data rates may require a hardware-based solution, our storage system will ensure that any off-the-shelf client can use the file system and decrypt files even if it is not equipped with custom hardware.

Security issues also become more interesting when wide-area access is taken into account. Currently, different security realms (e.g. Kerberos) are used at the DP laboratories, and unifying these into a globally accessible file system will be challenging. Our research group will work with our collaborators at the DP laboratories to address the issues involved in providing secure wide-area file system access.

2.2.6 Hardware

We plan to implement a proof-of-concept prototype of the storage system. This testbed implementation will serve as a development and demonstration platform, and will allow for implementation of prototype storage system components and testing of the algorithms and techniques under appropriately scaled down workloads. It will demonstrate the soundness of the basic techniques that will be developed under this research project. The prototype system will be constructed entirely from off-the-shelf components readily available from commercial vendors.

The prototype system will include:

- **Metadata servers** – two high-performance PC systems acting as either cooperating or redundant repositories of storage metadata. Depending upon the expected metadata workloads, these may incorporate high-performance RAID storage arrays for higher throughput and greater reliability.
- **OBSD systems** – eight high-performance PC systems acting as OBSDs. These will incorporate high-performance RAID storage arrays to achieve the desired throughput, storage capacity, and reliability.
- **Workload Drivers –** several additional high-performance PC systems dedicated to running test workloads and customized benchmarks against the storage system. The main requirement for these systems is the ability to generate a sufficiently demanding workload to effectively test the storage system. The functionality of these PC systems is limited to generating the workload, and accepting delivery of retrieved data. In order to increase perceived workload, it may be necessary to use each system to generate workloads equivalent to multiple real clients. Dedicated high-performance PCs should be able to do this effectively, but may require multiple NICs to actually send and receive the required volume of data across the network.
- **Network Infrastructure –** all computer systems will be equipped with Gigabit ethernet NICs. Additional network infrastructure components, probably Gigabit ethernet switches, will connect the prototype system components and will complete the testbed hardware. The number of ethernet switches will depend on the maximum achievable aggregate throughput of each switch. Enough bandwidth will be required to allow driving the OBSDs at their limits.

When the system is working well on our prototype, we will need to test scalability beyond the relatively small system we are building in our lab. We have two options for this: simulation and additional hardware. We can certainly simulate performance, but this approach will not be as convincing as running the code on real hardware. The second approach, running the system on more hardware, can be done either by installing additional hardware at UC Santa Cruz or one of the national labs, or by using hardware already present at the national labs. For example, we might be able to test a prototype file system on a set of workstations during relatively idle time such as a weekend or holiday. This approach would interfere little with lab operation, but would allow us to test the scalability of our design at little additional cost.

2.2.7 Summary

The National Laboratories need extremely high-bandwidth access to a multi-petabyte storage system to allow them to conduct the research crucial to understanding stockpile stewardship and other important problems. We propose to investigate issues of scalability, performance, and security that must be explored before such a system can be built.

By characterizing storage usage patterns at the National Laboratories, we will gain insight into the workload that a next-generation storage system must handle, allowing us to make good design decisions in building our prototype. By separating the storage system into scalable OBSD "bricks" connected by a high-bandwidth interconnection network and a scalable metadata cluster, we can build a system that will deliver very high transfer rates for the ASCI applications in a reliable way that scales up with the computing ability of each new system. At the same time, the storage system can serve visualization systems requesting data at multiple gigabytes per second, as well as workstations making many accesses to small files.

Performance is not the only storage issue of importance to ASCI, however. Our prototype will improve file system reliability by localizing storage allocation decisions to OBSDs and allowing each OBSD to maintain consistency locally, avoiding time-consuming global consistency checks. Furthermore, inclusion of snapshot capabilities along with improved reliability will reduce the need for full file system backups. Security is also an important issue; our prototype will include security mechanisms that work across the National Laboratories and, optionally, encrypt the data on the disks for greater security.

The Storage System Research Center at the University of California Santa Cruz brings together researchers with a tremendous amount of research and development experience in storage systems, computing systems, high-performance I/O, system security, and performance analysis. Through our discussions with personnel at LLNL and LANL we have developed a solid understanding of the mission, goals and requirements of the multi-petabyte storage systems needed for the next-generation high-performance computing systems being developed by and for these labs. Our proposed solution will directly target these systems and will provide storage architectures, techniques, and software that will address the storage capacity, I/O performance, access, and security needs of these systems.

2.3 Goals, Objectives, and Deliverables

2.3.1 Goals and Objectives

The primary research objective of the proposed work is the development of storage system techniques, algorithms, and architectures for high-performance multi-petabyte storage systems. Our research will be aimed at aiding the vendors of the post-Purple system in building the file system that can support the data rates necessary for the mission of the DP Laboratories.

Secondary goals of the work include investigating:

- A distributed file system model appropriate for supporting the high data rates, large file sizes, and high degree of parallelism inherent in terascale computing systems.
- Distributed storage system techniques capable of supporting the raw performance parameters for these systems including write performance, read performance, file creation performance, and shared-write performance.
- Robust distributed storage system techniques appropriate for the reliable medium- to long-term storage of multiple petabytes of data that cannot in general be backed up.
- Efficient, reliable, and robust Metadata Server techniques suitable for use in such a system.
- Efficient, reliable, and robust object-based storage device techniques suitable for use in such a system.
- Appropriate access control mechanisms that provide expected levels security while allowing efficient and appropriate access to data.
- Interconnection network network architectures and techniques capable of supporting expected throughputs of 10-100 GB/sec.

2.3.2 Deliverables

The deliverables of this research project will consist of actual work output and written reports. Written reports will include annual reports describing the results of the previous year's research, interim technical reports written as necessary to describe significant progress and results as they occur, and published papers describing specific research results and accomplishments as the research progresses. We will also deliver all models, traces, software, and reports generated under this project as well as the research prototype built to assist in the development and analysis of the ideas outlined in this proposal.

2.4 Key Tasks

Seven primary tasks are anticipated for the proposed work:

1. **Tri-Lab Interaction/Integration Plan**
   Coordinate and facilitate Tri-lab interaction and integration. Task Leader: Darrell Long.
2. **Workload Characterization**
   Characterize the workload anticipated for post-Purple storage systems based on analysis of existing workloads and projected workloads of next-generation systems. The output of this task will drive many aspects of the other tasks. Task Leader: Darrell Long.
3. **Metadata Server Software**
   Research efficient, reliable and robust software for clustered metadata servers including possible naming schemes, access control mechanisms, and metadata management and storage algorithms and techniques. Task Leader: Darrell Long.

4. **OBSD Software**
   Research efficient, reliable, and robust software for high-performance distributed object-based storage devices, including OBSD interface, allocation and scheduling algorithms. Task Leader: Ethan Miller.
5. **Client File System Software**
   Research client file system interfaces to high-performance distributed Metadata Servers and OBSDs. Task Leader: Scott Brandt.
6. **Interconnection Network**
   Research robust high-performance interconnection network techniques and architectures capable of meeting the throughput and availability needs of the post-Purple computing systems. Task Leader: Katia Obraczka.
7. **Analysis**
   Analyze and characterize the performance of the storage system design and proof-of-concept implementation. Task Leader: Alexandre Brandwajn.

At the present time our primary technical contact is Steve Louis, LLNL ASCI PSE Program Element Lead and Acting LLNL PSE Data Transfer and Storage Principal Investigator. If and when the project is funded we will work with him to identify appropriate contacts at each of the labs and for the individual tasks.

2.5 ASCI Relevance

High bandwidth storage systems are crucial to the ASCI project, as the following excerpts indicate:

> "Scalable Input/Output (I/O) for ASCI applications is an indispensable part of the infrastructure needed to support the high-level ASCI goals in weapons science, theory, and experimentation. Unless I/O transfer rates are in balance with the increasing computation speeds, ASCI applications will spend far too much time storing and retrieving data."
>
> "ASCI must develop flexible operating systems, I/O libraries, and database and information management systems, that can match such large-scale application I/O demands to storage hardware realities and visualization needs."
>
> "I/O and data handling remain fundamental bottlenecks in high-performance computing. The requirement for ASCI Scalable I/O is to deliver very high transfer rates for the ASCI applications in a reliable way that scales up with the computing capability of each new system."

The project described in this proposal, techniques for and a prototype of a high-performance multi-petabyte storage system, will be directly targeted for use in post-Purple computing systems planned for LLNL, LANL, and SNL. The research will be aimed at aiding the vendors of the post-Purple system in building a storage system that can support the data rates necessary for the mission of the DP Laboratories. Through a combination of a highly decentralized file system, high-performance algorithms, redundant storage, and parallelism, this system will directly address the storage capacity, I/O performance, access, and security needs of these systems.

Our specific research goal is to develop storage systems structures and algorithms that will meet the required I/O and storage system performance needs of the post-Purple computer. This research will address a key problem, as described above, that must be solved to enable the computations required for the stockpile stewardship mission of ASCI.

2.6 Interaction Plans

Several of the investigators have already established contacts with groups at LLNL. We have an ISCR grant to do basic research into emerging storage technologies (probe-based storage) for scientific computing. We also have just begun an ASCI level 3 collaboration on using object-based storage devices to address storage needs of the post-Purple computing system.

We propose to have quarterly meeting with our closest partners, the researchers and staff at LLNL. We also propose to to meet regularly with researchers and staff members at LANL and Sandia via travel to New Mexico and/or teleconference. We can schedule additional meetings as necessary; travel to LLNL is relatively easy for us because of our proximity.

We feel that early and regular contacts are important for us to fully understand the problems faced by the ASCI computing environment which is beyond anything seen anywhere else. If appropriate, the investigators that are United States citizens are willing to acquire a "Q" clearance if this will allow us to have discussions that enable us to fully address the problems that the National Laboratories face.

References

[1] A. Amer and D. D. E. Long, "Noah: Low-cost file access prediction through pairs," in *Proceedings of the IEEE International Performance, Computing and Communications Conference (IPCCC 2001)*, IEEE, Apr. 2001.

[2] R. C. Burns and D. D. E. Long, "Efficient distributed back-up with delta compression," in *Proceedings of I/O in Parallel and Distributed Systems (IOPADS)*, (San Jose), pp. 26–36, ACM, Nov. 1997.

[3] R. C. Burns and D. D. E. Long, "In-place reconstruction of delta compressed files," in *Proceedings of Principles of Distributed Computing (PODC)*, (Puerto Vallarta), pp. 267–75, ACM, June 1998.

[4] R. C. Burns, R. M. Rees, and D. D. E. Long, "Safe caching in a distributed file system for network attached storage," in *Proceedings of the International Parallel and Distributed Processing Symposium(IPDPS)*, (Cancun), pp. 155–162, IEEE, May 2000.

[5] R. C. Burns, R. M. Rees, L. Stockmeyer and D. D. E. Long, "Scalable session locking for a distributed file system," *Cluster Computing Journal*, 2001. to appear.

[6] W. Freeman and E. Miller, "An experimental analysis of cryptographic overhead in performance-critical systems," in *Proceedings of the 7th International Symposium on Modeling Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '99)*, (College Park, MD), pp. 348–57, Oct. 1999.

[7] W. Freeman and E. Miller, "Design for a decentralized security system for network attached storage," in *Proceedings of the 8th NASA Goddard Conference on Mass Storage Systems and Technologies / 17th IEEE Symposium on Mass Storage Systems*, pp. 361–373, 2000.

[8] G. A. Gibson, D. Nagle, K. Amiri, F. W. Chang, E. M. Feinberg, H. Gobioff, C. Lee, B. Ozceri, E. Riedel, D. Rochberg, and J. Zelenka, "File server scaling with network-attached secure disks," in *Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems*, pp. 272–284, 1997.

[9] R. Golding and D. D. E. Long, "Accessing replicated data in a large-scale distributed system," *International Journal in Computer Simulation*, vol. 1, no. 4, 1991.

[10] T. R. Haining, J.-F. Pâris, and D. D. E. Long, "A stack-based method for non-volatile cache management," in *Proceedings of the Eighth NASA Goddard Mass Storage Conference*, pp. 217–23, IEEE, March 2000.

[11] R. L. Haskin, "Tiger Shark – a scalable file system for multimedia," *IBM Journal of Research and Development*, vol. 42, pp. 185–97, Mar. 1998.

[12] T. M. Kroeger and D. D. E. Long, "Predicting future file-system actions from prior events," in *Proceedings of the 1996 Usenix Winter Technical Conference*, (San Diego), pp. 319–328, USENIX, Jan. 1996.

[13] T. M. Kroeger and D. D. E. Long, "The case for efficient file access pattern modeling," in *Proceedings of the Seventh Workshop on Hot Topics in Operating Systems (HotOS-VII)*, (Rio Rico, Arizona), IEEE, Mar. 1999.

[14] D. D. E. Long, "Application of probe-based storage to high-performance computing." Research proposal to Lawrence Livermore National Labs, Mar. 2000.

[15] E. Miller, D. Long, W. F. and B. Reed, "Strong security for distributed file system," in *20th IEEE International Performance, Computing, and Communications Conference(IPCCC 2001)*, (Phoenix, AZ), Apr. 2001.

[16] E. L. Miller and R. H. Katz, "RAMA: An easy-to-use, high-performance parallel file system," *Parallel Computing*, vol. 23, pp. 419–446, July 1997.

[17] K. Obraczka, P. Danzig, D. DeLucia, and E. Y. Tsai, "A tool for massively replicating internet archives: design, implementation, and experience," in *Proceedings of the Sixteenth International Conference on Distributed Computing Systems*, May 1996.

[18] J.-F. Pâris and D. D. E. Long, "Protecting replicated objects against media failures," in *Second International Workshop on Research Issues on Data Engineering: Transaction and Query Processing (RIDE-TQP'92)*, (Tempe, Arizona), pp. 109–15, Feb. 1992.

[19] A. Sweeney, D. Doucette, W. Hu, C. Anderson, M. Nishimoto, and G. Peck, "Scalability in the XFS file system," in *Proceedings of the 1996 USENIX Technical Conference*, (San Diego, CA), Jan. 1996.

[20] K. Wood, "Built to scale: The Hitachi Freedom Storage Lightning 9900 series," tech. rep., Hitachi Data Systems, 2000.

3 Institutional Capabilities

The Jack Baskin School of Engineering is a growing and focused engineering program, emphasizing computer science, computer engineering and electrical engineering, plus new interdisciplinary areas of computational biology, applied mathematics and statistics, and information systems management. Currently there are approximately 50 faculty in the school, and over 900 undergraduates and 160 graduate students, and the school has plans to more than double this number of faculty and students over the next ten years. A new engineering building is now being designed to house this growing set of academic programs and research. Areas of research excellence in the Jack Baskin School of Enginering include visualization and computer graphics, computer networks, computer and database systems, communications, VLSI and packaging, theoretical computer science, and bioinformatics.

The Center for Storage Systems Research brings together researchers whose focus has been on research into storage systems, computer system architecture, interconnection networks, performance evaluation, computer security, and related areas. The Center currently has six members and several affiliates and is growing rapidly. We have several open positions in the Computer Science Departments and also hope to hire faculty in the Electrical Engineering and Computer Engineering Departments. We are fortunate to have such a large group of faculty whose primary interest is storage system research, and to have a commitment from the University and the School of Engineering to grow this research area.

Within the SSRC we have extensive experience with techniques to improve the performance of storage systems. Our prior work has included predictive approaches to file system optimization [12, 13, 1], differential encoding for improved distributed data management [2, 3], and the use of NV-RAM for improved disk performance [10]. Our current research activities include work on automated data grouping aimed at improving the integration of mass storage systems into the storage hierarchy. This work investigates the automated grouping of related data items to improve storage systems performance. Through dynamic observation of data access patterns, we are able to construct groups of highly related data items. These groups can then be used to optimize data layout on mass storage systems, and when combined with caching algorithms can improve data retrieval performance for client systems. The results of this work are of direct utility to multiple domains, but for current storage architectures we perceive this to be of greatest benefit in improving the management of data stored on very large tape/disk silos, reducing the amount of high-latency operations incurred when using such a mass-store.

As traditional longitudinal magnetic recording approaches its hard physical limit, it will become necessary to look to other storage technologies for continued improvements in performance, capacity, and reliability. One area of SSRC research is the analysis and design of improved data management techniques in light of changing storage technologies. This places us in a better position to develop high performance systems using the best available technologies, while being well prepared to fully utilize upcoming storage trends and developments.

One technology showing great promise is orthogonal magnetic recording, employed in probe-based devices, fabricated with MEMS technology. These devices promise dramatic improvements in capacity, storage density, and overall performance. Although the adaptation of such devices to fit existing disk-based storage schemes is well underway, these devices are fundamentally different from conventional storage media. We are currently studying new paradigms made possible with these devices. Such an investigation at this stage of their development has the highest chance of influencing the products finally released by industry. In particular, we are already investigating these issues in light of workloads derived from ASCI applications and requirements provided by LLNL. This work [14], including the device simulator that has been developed, and our experience with the LLNL workloads is of direct benefit to the work we propose.

3.1 Investigator Qualifications

Dr. Darrell Long is Professor of Computer Science at UC Santa Cruz and Director of the Center for Storage Systems Research. For three years he served as Associate Dean of the Jack Baskin School of Engineering before becoming the Director of the Center. Dr. Long has 15 years experience in storage systems and is the author of more than 80 publications in the area of computer systems, most of which are in storage and I/O systems. He holds four patents and has an additional six pending, all in storage systems. He has been a consultant to IBM Research for six years, and is one of the architects of the IBM Storage Tank system, a very large-scale distributed direct-access storage system.

Dr. Scott Brandt is Assistant Professor of Computer Science at UC Santa Cruz. Dr. Brandt came to UC Santa after spending many years in industrial research and development at Honeywell Systems and Research Center, Alliant TechSystems Research and Technology Center, Secure Computing Corporation, and Theseus Research, a company he cofounded to resarch and develop asynchronous circuit technology. In all, he has 17 years of academic and industrial research and development experience in computer systems, computer architecture, operating systems, secure operating systems, and real-time image processing systems. He has 25 publications and six patents.

Dr. Ethan Miller is Assistant Professor of Computer Science at UC Santa Cruz. Dr. Miller spent his graduate career at the University of California at Berkeley, where he was a member of the RAID project. He came to Santa Cruz after spending six years on the faculty at the University of Maryland, where he was involved in several storage-related research projects. His

research has been sponsored by NASA and agencies in the Department of Defense, both organizations with very large storage needs, giving him significant experience in large-scale storage problems. He has ten years of experience and 40 publications, most in various aspects of storage systems.

Dr. Patrick Mantey is the Jack Baskin Professor of Computer Engineering at UC Santa Cruz and Director of the Organized Research Unit (ORU) that includes the Center for Storage Systems Research. Dr. Mantey was the founding Dean of the Jack Baskin School of Engineering and has extensive experience with computer systems both at IBM and in academia. He came to Santa Cruz in 1985 after a distinguished career at IBM Research, most of it at the Almaden Research Laboratory where IBM conducts the majority of its storage and database system research. He is a Co-PI on the NSF-supported UCSC Visualization Laboratory, and was a member of the 1988 NSF panel on Visualization in Scientific Computing.

Dr. Alexandre Brandwajn is Professor of Computer Engineering at UC Santa Cruz. Dr. Brandwajn is an expert in performance evaluation, and the author of numerous publications in the area of storage system performance. He came to Santa Cruz in 1985 following a distinguished career in academia and at Amdahl where his focus was on I/O system performance. He has extensive consulting experience in I/O systems, including work for Amdahl, Comparex and Storage Technology Corporation.

Dr. Katia Obraczka is Assistant Professor of Computer Engineering at UC Santa Cruz. Computer networks and distributed systems are the focus of her research. Before joining UCSC in January 2001, Dr. Obraczka was a research scientist at USC's Information Sciences Institute (ISI) and a research assistant professor at USC's Computer Science Department. Dr. Obraczka's current research interests include network protocol design and experimental evaluation, Internet information systems, Internet replication and caching, Internet measurement, and wireless mobile networking. One of Dr. Obraczka's current research projects (in collaboration with Argonne National Laboratories) is investigating making high-performance distributed computing applications network-topology-aware.

3.2 Computing Requirements

As part of conducting this research, we will construct a research prototype of the system to evaluate algorithms and techniques that we develop. The minimum configuration for such a prototype storage system is two metadata servers and eight OBSDs. In addition, the system will need appropriate interconnection hardware and sufficient client machines to exercise the storage system.

Our storage system prototype will need a minimum of eight OBSDs to test the scaling necessary for a petabyte-scale storage system. Some of the techniques we will use may require at least eight nodes; for example, ensuring redundancy in the face of a transient OBSD failure will require several nodes onto which the load of the failed node can be redirected. Moreover, it is impossible to make scaling claims or properly debug a large-scale storage system on any fewer than eight nodes. It is likely that we will need to perform larger-scale tests on our system to ensure further scalability, but we may be able to use existing systems at the national labs to do infrequent scalability tests. If our prototype is extensively tested on the small-scale system at UC Santa Cruz, however, we expect to be able to complete larger tests with minimum interference with lab activity.

We will build clustered metadata servers, starting with two nodes in the cluster and then adding nodes as appropriate to study the effects of scaling on these servers. While performing experiments on scalability of metadata service, we can retarget storage nodes to be part of the metadata cluster. This is possible because the handling of metadata requests and OBSD requests for data scale separately.

We will require a cluster of compute servers to provide clients to the storage system. We hope to obtain unclassified codes from our research contacts at the National Laboratories which can then be used to drive the storage system. We also have contacts at NASA and may be able to obtain codes that present challenges similar to those faced by the ASCI program.

3.3 Institutional Facilities

There are several activities in the School of Engineering that provide synergy with the proposed research. In particular, Prof. David Haussler has recently been named Howard Hughes Medical Investigator for his work on the Human Genome. As a result, Prof. Haussler's bioinformatics group is building a cluster of 1000 processors to be used in their gene finding and protein folding research. While this cluster is not large by National Laboratory standards, it is very large by academic research standards and provides us with a unique opportunity for collaboration. The codes run by the bioinformatics group are highly parallel, and we have explored with Prof. Haussler the idea of high performance search through vast collections of data.

The Visualization Laboratory is a joint venture among the School of Engineering and the Division of Natural Sciences. At the center of this laboratory is a large SGI Origin, which is used both by experts in visualization in the School of Engineering, but also by scientists in several departments in the Division of Natural Sciences. Recently there has been significant interest in moving from a single large visualization engine such the the SGI Origin to clusters built from commodity components. The existence of the Visualization Laboratory and our relationship with its investigators provides us with expertise in the needs of high-end visualization users, and also with possible clients for our prototype storage system.

The Center for Storage Systems Research has two servers with 1.2 TB of storage (small by National Laboratory standards), and two dual-processor 1 GHz computer severs with 4 GB of memory each. These systems can be used to store trace data, and to perform memory-intensive simulations of the various algorithms that we develop. We currently have a large collection of I/O system traces from industrial and an academic collaborators, and some traces from LLNL as part of our ISCR collaboration.

3.4 Leveraged Activities

At the University of California, Santa Cruz, we have a unique collection of talent within the Storage Systems Research Center (SSRC), and subsequently an excellent opportunity to leverage ongoing research activities for ASCI needs. Within our research center, we have a strong emphasis on systems issues involving the management of data and the effective use of storage systems. Of particular relevance is our ongoing research on high performance and scalable file systems, including several projects dealing with upcoming storage technologies that are both alternative and augumentative to traditional magnetic disk media.

The SSRC currently has several active research projects that leverage and are leveraged by the proposed research.

Currently funded research activities include:

- "Building a High-Performance Storage System from Commodity Components," Lawrence Livermore National Laboratory (ASCI Level 3 Collaboration). This project is an investigation of the use of object-based storage devices (OBSDs) in scalable storage systems. It is a preliminary investigation of the basic techniques and algorithms presented in this proposal.
- "Application of Probe-based Storage to High Performance Computing," Lawrence Livermore National Laboratory. This project is an exploration of the use of probe-based storage devices in file systems used by terascale computing.
- "Architectures and Algorithms to Exploit Probe-Based Storage," National Science Foundation. This project involves research on better models for probe-based storage, and measuring the performance of improved data placement heuristics and seek algorithms using these models.
- "High Performance Integration of Advanced Tertiary Stores," National Science Foundation. The goals of this project are to develop algorithms for grouping files based on usage so they can be moved to and from tertiary storage in large transfers.

4 Cost Estimate

4.1 Budget Justification

The majority of our budget for all three years is devoted to salaries for the investigators and graduate research assistants and related expenses. We have included 10 trips per year so that we may attend professional conferences to present our work and visit the National Labs to discuss our research.

The equipment budget in the first year is $50,000, allowing us to purchase systems on which to implement the prototype—OBSDs, metadata servers, interconnection hardware, and workload generators, as described in Section 2.2.6. We have allocated $25,000 per year in the successive years of the proposed research to purchase additional hardware to test the scalability of our prototype as well as upgrade previously purchased systems as necessary.

4.2 Leveraged Activities

- "HeRMES: High-performance Reliable MRAM-Enabled Storage," National Science Foundation (proposed). This research focuses on the construction of file systems using Magnetic RAM (MRAM) to dramatically improve performance and reliability. While MRAM is not currently planned for use in the terascale storage system, the HeRMES research is highly appropriate for use in such a system and, because it addresses storage system performance issues, will feed directly into the proposed work.
- "Uncertainty Quantification and Visualization of Computational Simulations," ASCI Level 2 (proposed). Members of the Visualization Laboratory and others at UC Santa Cruz are participating in an ASCI Level 2 proposal to investigate aspects of visualization in ASCI systems. While the two projects are distinct, we have direct and collegial communications between our research labs and the two proposed projects under the ASCI program will be highly complementary.

A Curriculum Vitae

A.1 Darrell D. E. Long

Employment History

2001–	Director, Storage Systems Research Center, University of California, Santa Cruz
1999–	Professor, Department of Computer Science, University of California, Santa Cruz
1995–	Visiting Scientist, IBM Almaden Research Center
1994–1999	Associate Professor, Department of Computer Science, University of California, Santa Cruz
1988–1994	Assistant Professor, Department of Computer Science, University of California, Santa Cruz

Education

Ph.D.	1988	University of California, San Diego, Computer Science and Engineering
M.S.	1986	University of California, San Diego, Computer Science
B.S.	1984	San Diego State University, Computer Science

Patents

1. Darrell D. E. Long and Thomas M. Kroeger. "Predictive Event Tracking Method," U.S. Patent No. 5,889,993, March 1999
2. Luis-Felipe Cabrera and Darrell D. E. Long. "Method and Apparatus for Establishing and Maintaining the Status of Membership Sets Used in Mirrored Read and Write Input/Output without Logging," U.S. Patent No. 5,917,998, June 1999
3. Randal C. Burns, Edward G. Chron, Darrell D. E. Long and Benjamin C. Reed. "Secure Array of Remotely Encrypted Storage Devices," U.S. Patent No. 5,931,947, August 1999
4. Randal C. Burns and Darrell D. E. Long. "A Method for Generating and Reconstructing In-place Delta Files," U.S. Patent No. 6,018,747, January 2000

Selected Publications

1. Randal C. Burns, Robert M. Rees, Larry Stockmeyer, and Darrell D. E. Long. "Scalable Session Locking for a Distributed File System," *Cluster Computing Journal*, 2001, *to appear*
2. Ethan L. Miller, Scott A. Brandt and Darrell D. E. Long. "HeRMES: High-Performance Reliable MRAM-Enabled Storage," *Proceedings of the Eighth Hot Topics in Operating Systems* (HotOS-VIII), Schloss Elmau, Germany: IEEE, May 2001
3. Randal C. Burns and Darrell D. E. Long. "An Analytical Study of Opportunistic Lease Renewal," *Proceedings of the Twenty-First International Conference on Distributed Computing Systems*, Phoenix: IEEE, April 2001
4. Randal C. Burns, Robert M. Rees and Darrell D. E. Long. "Safe Caching in a Distributed File System for Network Attached Storage," *Proceedings of the International Parallel and Distributed Processing Symposium* (IPDPS), Cancun: IEEE, May 2000
5. Thomas M. Kroeger and Darrell D. E. Long. "The Case for Efficient File Access Pattern Modeling," *Proceedings of the Seventh Hot Topics in Operating Systems* (HotOS-VII), Rio Rico, Arizona: IEEE, March 1999

A.2 Scott A. Brandt

Employment History

1999–	Assistant Professor, Department of Computer Science, University of California, Santa Cruz
1994–1999	Instructor, Teaching Assistant, Research Assistant, and Ph.D. student, Department of Computer Science, University of Colorado at Boulder
1992–1994	Senior Computer Scientist, Secure Computing Corporation, Roseville, Minnesota
1991–1992	Senior Research Scientist, Alliant Techsystems Research and Technology Center (Formerly Honeywell Systems and Research Center), Hopkins, Minnesota
1990–1991	Founder and Vice President, Theseus Research Inc., Minneapolis, Minnesota
1987–1990	Senior Research Scientist, Honeywell Systems and Research Center, Minneapolis, Minnesota

Education

Ph.D.	1999	University of Colorado, Boulder, Computer Science
M.S.	1993	University of Minnesota, Computer Science
B.S.	1987	University of Minnesota, Mathematics

Patents

1. K. Fant and S. Brandt. "Null Convention Logic System," U.S. Patent No. 5,828,228, October 1998
2. K. Fant and S. Brandt. "Method and System for Process Expression and Resolution," U.S. Patent No. 5,805,461, September 1998
3. K. Fant and S. Brandt. "Null Convention Logic System," U.S. Patent No. 5,664,212, September 1997
4. K. Fant and S. Brandt. "Method and System for Process Expression and Resolution Including a General Method of Direct Association," U.S. Patent No. 5,572,732, November 1996
5. K. Fant and S. Brandt. "Method and System for Process Expression and Resolution Including a Generally and Inherently Concurrent Computer Language," U.S. Patent No. 5,355,496, October 1994
6. K. Fant and S. Brandt. "Null Convention Logic System," U.S. Patent No. 5,305,463, April 1994

Selected Publications

1. Ethan L. Miller, Scott A. Brandt and Darrell D. E. Long. "HeRMES: High-Performance Reliable MRAM-Enabled Storage," *Proceedings of the 8th IEEE Workshop on Hot Topics in Operating Systems (HotOS-VIII)*, Schloss Elmau, Germany: IEEE, May 2001
2. S. Brandt and G. Nutt, "Flexible Soft Real-Time Processing in Middleware," *Real-Time Systems Journal, Special Issue on Flexible Scheduling in Real-Time Systems, to appear*
3. S. Brandt, "Performance Analysis of Dynamic Soft Real-Time Systems,"*Proceedings of the 20th IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pp. 379-386, April 2001
4. G. Nutt, S. Brandt, A. Griff, S. Siewert, T.Berk, and M. Humphrey, "Dynamically Negotiated Resource Management for Data Intensive Application Suites,"*IEEE Transactions on Knowledge and Data Engineering*, 12(1): 78–95, January/February 2000
5. S. Brandt, G. Nutt, T. Berk, and J. Mankovich, "A Dynamic Quality of Service Middleware Agent for Mediating Application Resource Usage," *Proceedings of the 19th IEEE Real-Time Systems Symposium (RTSS '98)*, pp. 307–317, December 1998

A.3 Ethan L. Miller

Employment History

2000–	Assistant Professor, Department of Computer Science, University of California, Santa Cruz
1999	System Architect, Endeca, Cambridge, MA
1994–2000	Assistant Professor, University of Maryland Baltimore County
1989–1993	Research Assistant (graduate student), RAID-II disk array project, University of California at Berkeley

Education

Ph.D.	1995	University of California, Berkeley, Computer Science
M.S.	1990	University of California, Berkeley, Computer Science
Sc.B.	1987	Brown University, Computer Science, *magna cum laude*

Honors

1989–1993	National Science Foundation Fellowship, University of California at Berkeley
1988–1989	Cal MICRO Fellowship, University of California at Berkeley
1987	William Gaston Prize for Academic Excellence (awarded to the top student in the department)

Professional Activities

1. Member of the IEEE Technical Committee on Mass Storage Systems Executive Committee, 1996–
2. Chair, IEEE Technical Committee on Operating Systems, 2001–
3. Program Committee Member, IEEE Mass Storage Systems Symposium, 1998–
4. Program Committee Member, USENIX Conference on File and Storage Technologies, 2002
5. Program Committee Member, Workshop on I/O in Parallel and Distributed Systems (IOPADS), 1996
6. Member of the ACM, IEEE, USENIX, and Sigma Xi

Selected Publications

1. Ethan Miller, Darrell Long, William Freeman, and Ben Reed, "Strong Security for Distributed File Systems," *20th IEEE International Performance, Computing, and Communications Conference (IPCCC 2001)*, Phoenix, AZ, April 2001
2. William Freeman and Ethan Miller, "Design for A Decentralized Security System For Network Attached Storage," *Proceedings of the 8th Goddard Conference on Mass Storage Systems and Technologies / 17th IEEE Symposium on Mass Storage Systems*, College Park, MD, March 2000
3. Ethan L. Miller, Dan Shen, Junli Liu, and Charles Nicholas, "Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System," *Journal of Digital Information* (online refereed journal), January 2000
4. Theodore Johnson and Ethan Miller, "Performance Measurements and Models of Tertiary Storage Devices," *Proceedings of the 1998 Conference on Very Large Databases* (VLDB '98), New York, NY, August 1998
5. Ethan Miller and Randy H. Katz, "RAMA: An Easy-To-Use, High-Performance Parallel File System," *Parallel Computing*, **23**(4), July 1997

A.4 Patrick E. Mantey

Employment History

1997–2001	Dean, Jack Baskin School of Engineering, University of California, Santa Cruz
1984–1997	Professor and Chair, Computer Engineering, University of California, Santa Cruz
1967–1984	IBM Research Division, Manager, Engineering and Special Systems Department, Computer Science
1978–1979	Visiting Associate Professor, Electrical Engineering; Lecturer, Graduate School of Business, Stanford University

Education

Ph.D.	1965	Stanford University, Electrical Engineering
M.S.	1961	University of Wisconsin, Electrical Engineering
B.A.	1960	*magna cum laude*, University of Notre Dame, Electrical Engineering

Recent Research Support

1997–2000	"Acquisition of UCSC Scientific Visualization Laboratory", National Science Foundation, Major Research Instrumentation Program(MRI), $460,000
1997–2002	"Interactive Visualization of Real-Time Databases", National Science Foundation, National Partnership for Advanced Computing Infrastructure (NPACI), $830,000

Professional Activities

1. Conference Committee for ACM MultiMedia, 1993 and 1994
2. Co-chairman of the program committee of the First International Conference on Computer Workstations, sponsored by the IEEE Computer Society, held in San Jose in November, 1985; and Conference Co-chairman for the Second IEEE Workstation Conference, held in Santa Clara in March 1988
3. Member of the NSF panel on Visualization in Scientific Computing, 1988
4. Board of Directors of the NSF-sponsored National Center of Geographic Information and Analysis, 1988–1992
5. Fellow of Institute of Electrical and Electronics Engineering and a member of ACM

Selected Publications

1. Darrell Long, Patrick Mantey, Eric Rosen, Craig Wittenbrink and Bruce Gritton. "REINAS: A Real-Time System for Managing Environmental Data," *Journal of Software Engineering and Knowledge Engineering*, vol. 8, no. 1 (March 1998), pp. 35–53
2. Eric Rosen, Theodore Haining, Darrell Long, Patrick Mantey and Craig Wittenbrink. "REINAS: A Real-Time System for Managing Environmental Data." *Proceedings of the Eighth Software Engineering and Knowledge Engineering Conference*, pp. 293–300, Lake Tahoe: SEKE, June 1996
3. Jehan-François Pâris, Darrell Long, And Patrick Mantey. "A Zero-Delay Broadcasting Protocol for Video on Demand," *Proceedings of the 1999 ACM International Multimedia Conference*, Orlando, Florida, October 1999
4. Craig Wittenbrink, Eric Rosen, Alex Pang, Suresh K Lodha, and Patrick Mantey. "Real-Time Database Support for Environmental Visualization," *Proceedings of the Second Workshop on Database Issues for Data Visualization*, Visualization '95, Atlanta, GA, 1995
5. David E. Levy and Patrick E. Mantey. "Electronic Libraries and Optical Jukebox Server Scheduling," *Proceedings of SPIE/IS&T Symposium on Electronic Imaging Science & Technology*, February 1992

A.5 Alexandre Brandwajn

Employment History

1985–	Professor, Department of Computer Engineering, University of California, Santa Cruz
1991–1993	Director of Education Abroad Program Centers in France (Bordeaux, Pau, Poitiers, Toulouse)
1982–1985	Manager, Systems Analysis Department, Amdahl Corporation
1979–1982	Senior Computer Architect, Performance Analysis Department, Amdahl Corporation, Sunnyvale, California
1978–1979	Professor of Computer Science, Ecole Nationale Supérieure des Télécommunications, Paris
1978–1979	Associate Professor of Computer Science, Duke University, North Carolina (sabbatical year)
1974–1977	Associate Professor of Computer Science, Ecole Nationale Supérieure des Télécommunications, Paris
1971–1975	Researcher-Operating System Design, Performance Evaluation, InRIA, Roquencourt, France

Education

1975	Docteur d'Etat (ès-Sciences Mathematiques), Computer Science, University of Paris VI
1973	Licencie es-Lettres (BA), Russian Literature, University of Paris IV Sorbonne
1972	Docteur-Ingenieur (Doctor of Engineering), Computer Science, University of Paris VI
1971	Ingenieur Civil des Télécommunications, Ecole Nationale Supérieure des Télécommunication, Paris

Consulting

1998–	Consultant to Storage Technology Corporation – I/O Subsystem Performance, Model Calibration
1992–1993	Consultant to Comparex Information Systems (Germany) – RAID-5 I/O Controller Performance
1985–1991	Consultant to Amdahl Corporation – I/O Subsystem Performance

Professional Activities

1. General Chair for the ACM SIGMETRICS 2000 Conference
2. Committee member for numerous conferences and symposia, including International ATM Conference, July 2000

Selected Publications

1. A. Brandwajn and Anil K. Sahai. "Aspects of the Solution of Some Multiclass Loss Systems," *Performance Evaluation*, vol. 17, 1993, pp. 141–159
2. A. Brandwajn and A. K. Sahai. "A Note on Approximate Iterative Solution of Open Tandem Networks with Blocking," *Performance Evaluation*, vol. 10, 1989, pp. 211–218
3. A. Brandwajn. "Load Imbalance in DASD Dynamic Reconnection," *Information Processing Letters*, vol. 28, 1988, pp. 111–119
4. A. Brandwajn. "A Finite-Source Multiserver Queue with Preemptive Priorities," *Proceedings of the Fourth International Workshop on Queueing Networks with Finite Capacity*, Ilkley, UK, July 2000
5. A. Brandwajn and D. Levy. "A Study of Cached RAID-5 I/O," *Computer Measurement Group Conference*, December 1994

A.6 Katia Obraczka

Employment History

2001–	Assistant Professor, Computer Engineering Department, University of California, Santa Cruz
1997–2001	Research Assistant Professor, Computer Science Department, University of Southern California
1995–2001	Research Scientist, Information Sciences Institute, University of Southern California
1994–1995	Manager of Networking and Distributed Systems Laboratory, University of Southern California
1990–1994	Research Assistant, Computer Science Department, University of Southern California
1992–1996	Instructor, Computer Science Department and Programming and Data Processing Department, University of Southern California
1989–1990	Teaching Assistant, Computer Science Department, University of Southern California

Education

Ph.D.	1994	University of Southern California, Computer Science
M.S.	1990	University of Southern California, Computer Science
M.S.	1987	Federal University of Rio de Janeiro, Computer Engineering
B.S.	1981	Federal University of Rio de Janeiro, Electrical Engineering

Professional Activities

1. Member of the Internet Research Task Force (IRTF) Reliable Multicast Working Group (RMWG)
2. Program Committee member: USENIX USITS '99, ACM/IEEE Mobihoc 2000, and IEEE ICDCS 2001
3. Conference Reviewer: ACM SIGCOMM and IEEE INFOCOMM, ACM SIGMETRICS, and IEEE ICDCS
4. Journal Reviewer: Journal of Internetworking, ACM/IEEE Transactions on Networking, IEEE Transactions on Parallel and Distributed Systems
5. Reviewer and panelist for NSF Networking and Database and Expert Systems Programs
6. ACM and IEEE Member

Selected Publications

1. K. Obraczka and P. Danzig. "Evaluating the Performance of Flood-d: A Tool for Efficiently Replicating Internet Information Services," *IEEE Journal on Selected Areas of Communication*, April 1998
2. K. Obraczka and G. Gheorghiu. "The Performance of a Service for Network-Aware Applications," *Proceedings of the ACM Sigmetrics SPDT '98*
3. K. Obraczka, P. Danzig, D. DeLucia, and E.-Y. Tsai. "A Tool for Massively Replicating Internet Archives: Design, Implementation and Experience," *Proceedings of the Sixteenth International Conference on Distributed Computing Systems*, May 1996
4. K. Obraczka, P. Danzig, and S.-H. Li. "Internet Resource Discovery Services," *IEEE Computer Magazine*, vol. 26, no. 5, September 1993, pp. 8–22
5. P. Danzig, K. Obraczka, and A. Kumar. "An Analysis of Wide-Area Name Server Traffic: A Study of the Domain Name System," *Proceedings of the ACM SIGCOMM '92*, August, 1992