LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-235259

Scalable File Systems for High Performance Computing Final Report

S. A. Brandt

October 4, 2007

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Scalable File Systems for High Performance Computing Final Report

Scott A. Brandt (Principle Investigator)
Computer Science Department
University of California, Santa Cruz

July 16, 2007

1 Executive Summary

1.1 Overview

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 research has successfully addressed a number of the key research issues in the design of a high-performance multi-petabyte storage system 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).

Our ASCI-sponsored research in addressing these problems has resulted in the most successful research project in the history of the UCSC Computer Systems Research Group. It has resulted in 35+ publications, 10+ Ph.D.s and M.S.s, numerous technical accomplishments including a variety of effective data and metadata management solutions and a complete working prototype of Ceph, petascale distributed high-performance data storage system, which has been released to the public under an open-source license.

Driven by the data rates and usage scenarios that characterize the mission of the Defense Programs (DP) Laboratories, our research has successfully addressed a number of the key research issues in the design of a high-performance multi-petabyte storage system 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). Our research addresses these issues and achieves the overall goal of developing 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 and other general-purpose access patterns. Our results have been validated and demonstrated by the creation of proof-of-concept implementations of this architecture culminating with the development and open-source release of Ceph. A major value of this research is its use by the DP Laboratories in providing direction to the vendors of post-Purple systems in building new file and storage systems.

1.2 Meeting the Challenge

The ASCI program provided unique challenges for researchers that are not found in any other environment. The post-Purple requirement of 10s–100sof PB (petabytes) of data delivered at a sustained 1 TB/s or more is beyond what was even contemplated today by most researchers in the past. Scalability is the key to achieving this goal. While storage components have continued to increase in density and in speed, they have not improved quickly enough to meet this goal without a high degree of parallelism. We have met this goal using a highly scalable architecture made up of commodity components.

Our storage system architecture (shown in Figure **??**) is based upon object-based storage devices (OSDs) connected by high-speed networks. A key advantage of OSDs 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 OSD a given segment should be placed. Since this decision is quite simple and allows massive parallelism, each

OSD need only manage concurrency locally, allowing a file system built from thousands of OSDs to achieve massively parallel data transfers. Additionally, OSDs 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 OSD-based file systems (we have developed two, OBFS and EBOFS) are based upon a flat name space, employing a cluster of metadata servers to translate "human-readable" names into a file identifier and, ultimately, object identifiers, in this flat namespace. OBFS and EBOFS both provide high-performance object storage with relatively little code, taking advantage of the flat namespace, lack of directories, and lack of inter-object locality to manage objects much more efficiently than is possible with typical file systems.

The use of a cluster metadata servers provides both high performance and redundancy, allowing the metadata service to scale with the number of file system clients. Our metadata management architecture, Dynamic Subtree Partitioning, dynamically redelegates and, as needed, replicates responsibility for subtrees of the metadata hierarchy among the metadata servers in the cluster to balance the workload, manage hotspots, and allow for automatic scaling when new metadata servers are added.

Scalability also requires the ability to add new OSDs (or remove old ones). To address this need we have developed RUSH and CRUSH. RUSH is a family of algorithms that allocate objects to servers. RUSH allows for good random allocation of objects (for load balancing) but, unlike simple hashing, allows new OSDs to be added to (or removed from) the system without all of the overhead usually required to rehash and relocated existing objects. RUSH also enables replication across multiple OSDs for redundancy and/or parallelism. CRUSH extends the RUSH algorithms and encapsulates them in a framework enabling automatic mirroring across user-specifiable failure domains.

Built with commodity hard drives, the raw number of drives required to store petabytes of data means that our storage system can be expected to have frequent drive failures. Our research has shown that traditional RAID cannot adequately protect the system from data loss, even in the short term. Our Fast Recovery Mechanisms (FaRMs) address this problem, rapidly reconstructing lost disks on an object-by-object basis and providing the high reliability required of such systems. Our Reliable Autonomic Distributed Object Store (RADOS) abstracts the notion of a single object store into a reliable distributed object store, using FaRM techniques to quickly recover from failed drives and to handle the addition of new storage.

Understanding the workloads the storage system must handle is critical to developing a system capable of meeting its performance requirements. While there is a significant body of research on general-purpose workloads, relatively little research discusses the specific workloads that will be encountered in post-Purple computing platforms. In addition, due to the lack of large object-based storage systems, nothing has previously been published on the object workload that results from even well-understood file- and block-level workloads. Our research in this area has resulted in a better understanding of the expected scientific object workloads and enabled both more informed system development and more accurate test workload generation.

Security is also an important design consideration for distributed file systems. We have developed OSD-specific security architectures and protocols that allow for secure and efficient operation within the system. One challenging aspect of the problem is that the client is the only portion of the system that may have a consistent view of the file-to-object mapping, but OSDs must verify all accesses to ensure that they have been granted by the metadata servers. We have addressed this problem. We have also developed a set of theoretical results that begin to allow reasoning about security in distributed storage systems.

Large distributed storage systems are shared by many users and applications. Existing storage systems cannot guarantee the performance required to capture simulation data or deliver visualization data while handling mixed workloads. The standard solution is to partition the use of the resources in time, with some uses prohibited while other high-priority uses are active, with overall inefficient (and inconvenient) use of an expensive shared resource. We have developed algorithms and solutions to provide Quality of service with direct-attached disks, object stores, and distributed object-based storage systems.

This project has also had solid educational benefits. Student researchers on this project have earned (or will soon earn) 5 Ph.D.s and 5 Master's degrees. Related to this research we have introduced new graduate courses in Storage Systems and Distributed Systems, and have developed numerous related course projects in these and other classes.

2 Technical Goals

The ASCI program has provided unique I/O challenges for researchers that are not found in any other environment. These include 10s–100s PB (petabytes) of data, sustained data rates of 1 TB/s, thousands of hard drives moving data at their maximum sustained rate, billions of files ranging from bytes to terabytes, hundreds of thousands of files per directory (or more), and a requirement for very low-latency metadata access.

Because of the demands imposed by the large amount of legacy code as well as existing interfaces, tools, and programmer skills, a POSIX-like interface is required, including standard file/directory semantics. To support the intended applications, parallel accesses must be supported from 100,000+ clients accessing different files in different directories, the same directory,

and even the same file. Wide-area general-purpose access is also required, and the ability to manage performance so that the system can guarantee performance under mixed workloads is highly desirable.

We have achieved these goals with a highly scalable architecture made up of commodity components. At the same time, we have achieved a number of other accomplishments crucial to the overall ASCI mission, as discussed in the following sections.

3 Summary of Accomplishments

Our ASCI-sponsored research has resulted in 7 key accomplishments in the area of scalable file systems for high-performance computing:

1. We have developed a range of tools and techniques for petascale I/O and data management.
2. Building upon (and embodying) these tools and techniques, we have developed Ceph, "the most advanced distributed storage system in the world" (from a member of the storage systems research community at OSDI '06).
3. We have helped publicize HEC I/O needs to the academic storage systems community
4. We have further developed and publicized the object-based storage model as a basis for achieving scalable, high-performance, distributed storage.
5. We have trained a cadre of researchers knowledgeable and concerned about HEC I/O needs
6. We have published 35+ research papers and theses related to our ASCI-funded high-performance I/O research
7. We have open-sourced the results of our efforts, showing how to implement the techniques we developed, providing a prototype storage system that can be used as the basis for further development, and providing a baseline against which other solutions can be compared.

4 Technical Achievements

In pursuing this research we have developed a range of tools and techniques for petascale I/O and data management, including:

- RUSH: Replication Under Scalable Hashing—a family of algorithms for pseudorandomly distributing data across a set of object storage devices with or without replication.
- CRUSH: Controlled Replication Under Scalable Hashing—a system for pseudorandomly distributing data across a set of object storage devices with or without replication, including code for managing device failure, addition and removal, and overload, with the ability to distribute replicas across failure domains according to an administrator-specified algorithm.
- FaRMs: Fast Recovery Mechanisms—techniques for robustly storing data across a set of object storage devices under device failures and quickly recovering from such failures.
- LH3: Lazy Hybrid Hierarchical Hashing—a distributed metadata management architecture based upon hashing.
- DSP: Dynamic Subtree Partitioning—a robust high-performance distributed metadata management architecture.
- OBFS: Object-Based File System—a high-performance local file system for object storage.
- EBOFS: Extent-based Object File System—a high-performance local file system for object-storage.
- RADOS: Robust Autonomic Distributed Object Storage—a distributed scheme for managing a set of object storage devices (each running a local file system for object storage) that provides the illusion of one large, reliable object store.
- Ceph: Our high-performance distributed object-based storage system. Includes DSP, CRUSH, RADOS, and EBOFS.
- Storage Quality of Service—Quality of service support for storage systems.
- AQuA: A Quality of Service Aware OSD—A modified version of OBFS that provides local storage Quality of Service according to user/administrator specifications.

- Bourbon: A QoS-aware version of Ceph, built upon QoS-aware versions of EBOFS, that provides system level distributed storage QoS according to user/administrator specifications.
- Security Algorithms: A set of security algorithms for use in distributed object-based storage systems.
- Security Theory: A basic body of theory allowing one to formally prove theorems about the security of distributed object-based storage systems.
- Workload Characterization: An analysis and characterization of ASCI-relevant high-performance I/O workloads and the resulting object workloads that they can be expected to generate.
- Networking: An analysis of some issues in networking for large, distributed storage systems.

5 People

A large number of UCSC faculty and students have been actively involved in this research project, including a large number of students who have been placed in and are now actively working in the storage industry.

Faculty

- Prof. Scott Brandt
- Prof. Darrell Long
- Assoc. Prof. Ethan Miller
- Prof. Martin Abadi

Postdoc/Research Faculty

- Dr. Carlos Maltzahn (formerly on leave from Network Appliance, Inc. and now Research Professor at UCSC)

Visiting/Affiliated Faculty

- Prof. Emilia Rosti

Ph.D. Students

- Dr. Chris Xin (now at Symantec)
- Dr. Feng Wang (now at Symantec)
- Dr. Joel Wu
- Sage Weil (graduating soon)
- Avik Chaudhury (graduating this coming year)

M.S. Students

- Lan Xue (now at VMware)
- R.J. Honicky (now at UC Berkeley)
- Chris Olson (now at Google)
- Andrew Leung (continuing on to Ph.D.)
- Martin Arnberg (next year)

Other contributers

- Dr. Zachary Peterson
- Dr. Bo Hong (now at Symantec)
- Dr. Scott Banachowski (now at Yahoo!)
- Dr. Timothy Bisson (now at Network Appliance)
- Kristal Pollock (now at IBM Almaden)
- David Bigelow (continuing Ph.D. student)
- Anna Povzner (continuing Ph.D. student)
- Tim Kaldewey (continuing Ph.D. student)
- Suresh Iyer (continuing Ph.D. student)
- Eric LaLonde (continuing M.S. student)
- Stephanie Jones (continuing M.S. student)

6 Degrees Awarded

Ph.D.s

- Dr. Chris Xin (2005)
  Dissertation: *Understanding and Coping with Failures in Large-Scale Storage Systems*
  Advisor: Prof. Ethan Miller
- Dr. Feng Wang (2006)
  Dissertation: *Storage Management in Large Distributed Object-based Storage Systems*
  Advisor: Prof. Scott Brandt
- Dr. Joel Wu (2007)
  Dissertation: *Providing Quality of Service Support for File Systems*
  Advisor: Prof. Scott Brandt
- Sage Weil (expected 2007)
  Dissertation: *Ceph: A Scalable, High-performance, Distributed Object-based File System*
  Advisor: Prof. Scott Brandt
- Avik Chaudhury (Expected early 2008)
  Dissertation: *Formal Security for Distributed File Systems*
  Advisor: Prof. Martin Abadi

M.S.s

- Lan Xue (2003)
  Project: *Efficient Metadata Management in Large Distributed File Systems*
  Advisor: Prof. Scott Brandt
- R.J. Honicky (2004)
  Project: *Object Placement Algorithms for OBSD Systems*
  Advisor: Prof. Ethan Miller

- Chris Olson (2005)
  Project: *Security in Ceph*
  Advisor: Prof. Ethan Miller

- Andrew Leung (expected 2007)
  Project: *Scalable Security for High-Performance Storage Systems*
  Advisor. Prof. Ethan Miller

- Martin Arnberg (Expected early 2008)
  Project: *Quotas in Ceph*
  UCSC advisor: Prof. Darrell Long

7 Publications

The following is a list of our publications most closely related to this research.

7.1 Data Distribution: RUSH and CRUSH

RUSH

- R. J. Honicky and Ethan L. Miller, "An Optimal Algorithm for Online Reorganization of Replicated Data," SSRC Technical Report UCSC-CRL-02-36, Storage Systems Research Center, University of California, Santa Cruz, Nov. 2002.

- R. J. Honicky, Ethan L. Miller, "A Fast Algorithm for Online Placement and Reorganization of Replicated Data," Proceedings of the 17th International Parallel and Distributed Processing Symposium (IPDPS 2003), April 2003

- R. J. Honicky, Ethan L. Miller, "Replication Under Scalable Hashing: A Family of Algorithms for Scalable Decentralized Data Distribution," Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS 2004), April 2004.

- R.J. Honicky, "Object Placement Algorithms for OBSD Systems," M.S. Project, Computer Science Department, University of California, Santa Cruz, June 2004.

CRUSH

- Sage A. Weil and Scott A. Brandt and Ethan L. Miller and Carlos Maltzahn, "CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data," SSRC Technical Report SSRC-06-01, Storage Systems Research Center, University of California, Santa Cruz, Jan, 2006.

- Sage Weil, Scott A. Brandt, Ethan L. Miller, Carlos Maltzahn, "CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data," Proceedings of SC '06, November 2006.

7.2 Metadata Management: LH and DSP

LH

- Scott A. Brandt, Lan Xue, Ethan L. Miller, Darrell D. E. Long, "Efficient Metadata Management in Large Distributed File Systems," Proceedings of the 20th IEEE / 11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST 2003), April 2003.

- Lan Xue, "Efficient Metadata Management in Large Distributed File Systems," M.S. Project, June 2003.

- Kristal Pollack, Scott A. Brandt, "Efficient Access Control For Distributed Hierarchical File Systems," IEEE / Nasa Goddard Conference On Mass Storage Systems And Technologies (MSST 2005), April 2005.

DSP

- Sage Weil, Scott A. Brandt, Ethan L. Miller, Kristal Pollack, ”Intelligent Metadata Management For A Petabyte-Scale File System,” 2nd Intelligent Storage Workshop, May 2004.
- Sage Weil, Kristal Pollack, Scott A. Brandt, Ethan L. Miller, ”Dynamic Metadata Management For Petabyte-Scale File Systems,” Proceedings Of The 2004 ACM/IEEE Conference On Supercomputing (SC ’04), November 2004.

7.3 Object Storage: OBFS, EBOFS, and RADOS

OBFS

- Feng Wang, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, ”OBFS: A File System for Object-Based Storage Devices,” Proceedings of the 21st IEEE / 12th NASA Goddard Conference on Mass Storage Systems and Technologies, April 2004, pages 283-300.
- Feng Wang, Storage Management in Large Distributed Object-Based Storage Systems, Ph.D. thesis, University of California, Santa Cruz, December 2006.

EBOFS

- Sage A. Weil, ”Maximizing OSD Performance with EBOFS,” SSRC Technical Report SSRC-04-02, Storage Systems Research Center, University of California, Santa Cruz, March, 2004.

RADOS

- Sage A. Weil, ”Scalable Archival Data and Metadata Management in Object-based File Systems,” SSRC Technical Report SSRC-04-01, Storage Systems Research Center, University of California, Santa Cruz, May, 2004.
- Sage Weil, Carlos Maltzahn and Scott A. Brandt, ”RADOS: A Reliable Autonomic Distributed Object Store,” SSRC Technical Report SSRC-07-01, Storage Systems Research Center, University of California, Santa Cruz, January, 2007.

7.4 Reliability

- Qin Xin, Ethan L. Miller, Thomas Schwarz, Darrell D. E. Long, Scott A. Brandt, Witold Litwin, Reliability Mechanisms for Very large Storage Systems, Proceedings of the 20th IEEE / 11th NASA Goddard Conference on Mass Storage Systems and Technologies, April 2003, pages 146-156.
- Qin Xin, Ethan L. Miller, Thomas Schwarz, Evaluation of Distributed Recovery in Large-Scale Storage Systems, Proceedings of the 13th IEEE International Symposium on High Performance Distributed Computing (HPDC 2004), June 2004, pages 172-181.
- Qin Xin, Thomas Schwarz, Ethan L. Miller, Disk Infant Mortality in Large Storage Systems, Proceedings of the 13th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS ’05), September 2005.
- Jehan-Franois Pris, Darrell D. E. Long, Using Device Diversity to Protect Data against Batch-Correlated Disk Failures, Proceedings of the 2nd ACM Workshop on Storage Security and Survivability (StorageSS 2006), October 2006.
- Qin Xin, Understanding and Coping with Failures in Large-Scale Storage Systems, Ph.D. thesis, Computer Science Department, University of California, Santa Cruz, June 2005 (Technical Report UCSC-SSRC-07-06).

7.5 Security Algorithms and Theory

Security Algorithms

- Ethan L. Miller, Darrell D. E. Long, William E. Freeman, Benjamin C. Reed, Strong Security for Network-Attached Storage, Proceedings of the 2002 Conference on File and Storage Technologies (FAST), January 2002, pages 1-13.
- Thomas Schwarz, Qin Xin, Ethan L. Miller, Darrell D. E. Long, Andy Hospodor, Spencer Ng, Disk Scrubbing in Large Archival Storage Systems, Proceedings of the 12th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '04), October 2004, pages 409-418.
- Chris Olson, "Security in Ceph," M.S. project, Computer Science Department, University of California, Santa Cruz, June 2005.
- Christopher Olson, Ethan L. Miller, Secure Capabilities for a Petabyte-Scale Object-Based Distributed File System, Proceedings of the 2005 ACM Workshop on Storage Security and Survivability (StorageSS 2005), November 2005.
- Andrew Leung, Ethan L. Miller, Scalable Security for Large, High Performance Storage Systems, Proceedings of the 2nd ACM Workshop on Storage Security and Survivability (StorageSS 2006), October 2006.

Security Theory

- Avik Chaudhuri and Martn Abadi, Formal Security Analysis of Basic Network- Attached Storage, Proceedings of the 3rd ACM Workshop on Formal Methods in Security Engineering (FMSE 2005), pp. 43-52, Nov. 2005.
- Avik Chaudhuri and Martn Abadi, Secrecy by Typing and File-Access Control, Proceedings of the 19th IEEE Computer Security Foundations Workshop, Venice, Italy (CSFW 2006), pp. 112-123, July, 2006.
- Avik Chaudhuri, Dynamic Access Control in a Concurrent Object Calculus, Proceedings of the 17th International Conference on Concurrency Theory, Bonn, Germany (CONCUR 2006), pp. 263-278, Aug. 2006.
- Avik Chaudhuri and Martn Abadi, Formal Analysis of Dynamic, Distributed File-System Access Controls, Proceedings of the 26th IFIP WG6.1 International Conference on Formal Methods for Networked and Distributed Systems, Paris, France (FORTE 2006), pp. 99-114, Sept. 2006.

7.6 Storage QoS, including AQuA and Bourbon

- Joel Wu and Scott A. Brandt, "QoS Support for Intelligent Storage Devices," Intelligent Storage Workshop, University of Minnesota Digital Technology Center, May 2004.
- Joel C. Wu and Scott A. Brandt, "Storage Access Support for Soft Real-Time Applications," the 10th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2004), May 2004.
- Joel Wu, Scott Banachowski and Scott A. Brandt, "Automated QoS Support for Multimedia Disk Access," Conference on Multimedia Computing and Networking, San Jose, CA, January 2005, pp. 103–107.
- Joel C. Wu and Scott A. Brandt, "Hierarchical Disk Sharing for Multimedia Systems and Servers," Proceedings of the 15th ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 2005), June 2005, pages 189–194.
- Joel C. Wu and Scott A. Brandt, "QoS Support in Object-Based Storage Devices," International Workshop on Storage Network Architecture and Parall I/Os (SNAPI '05), St. Louis, MO, September 2005, pp. 41–48.
- Joel C. Wu and Scott A. Brandt, "The Design and Implementation of AQuA: an Adaptive Quality of Service Aware Object-Based Storage Device," Proceedings of the 23rd IEEE / 14th NASA Goddard Conference on Mass Storage Systems and Technologies, May 2006, pp. 209–218.
- Joel C. Wu, Bo Hong, Scott A. Brandt, Ensuring Performance in Activity-Based File Relocation, Proceedings of the International Performance Conference on Computers and Communication (IPCCC '07), April 2007, pages 75–84.

- Dr. Joel Wu, "Providing Quality of Service Support for File Systems," Ph.D. Thesis, Computer Science Department, University of California, Santa Cruz, June 2007.
- Joel Wu and Scott A. Brandt, Providing Quality of Service Support in Object-based File Systems, IEEE / NASA Goddard Conference On Mass Storage Systems And Technologies (MSST 2007), to appear.

7.7 Workload, Networking, and Performance

Workload

- Feng Wang, Qin Xin, Bo Hong, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, Tyce T. Mclarty, File System Workload Analysis For Large Scientific Computing Applications, NASA/IEEE Conference On Mass Storage Systems And Technologies (MSST 2004), April 2004, Pages 139152.

Networking

- Andy Hospodor, Ethan L. Miller, Interconnection Architectures for Petabyte- Scale High-Performance Storage Systems, Proceedings of the 21st IEEE / 12th NASA Goddard Conference on Mass Storage Systems and Technologies, April 2004, pages 273-281.
- Qin Xin, Ethan L. Miller, Thomas Schwarz, Darrell D. E. Long, Impact of Failure on Interconnection Networks in Large Storage Systems, Proceedings of the 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies, April 2005.

Performance

- Andrew Leung, Eric Lalonde, Jacob Telleen, James Davis, Carlos Maltzahn, "Using Comprehensive Analysis for Performance Debugging in Distributed Storage Systems," Technical Report UCSC-SSRC-07-05, May 2007.

7.8 Ceph

- Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long and Carlos Maltzahn, "Ceph: A Scalable Object-based Storage System," SSRC Technical Report SSRC-06-02, Storage Systems Research Center, University of California, Santa Cruz, January 2006.
- Sage Weil, Scott A. Brandt, Ethan L. Miller, Darrell D. E. Long, Carlos Maltzahn, "Ceph: A Scalable, High-Performance Distributed File System," Proceedings of the 7th Conference on Operating Systems Design and Implementation (OSDI '06), November 2006.

8 Technology Transfer

This project has generated significant technology transfer through five primary vehicles:

- Publication—Through a large number of highly visible publications culminating with our the November 2006 publication of Ceph at OSDI, the most cited publication venue in Computer Science (according to Citeseer), our publications have generated significant interest in industry and gathered a large number of citations in academia.
- Placement of students—Graduates of this project have been placed in a number of companies concerned with large and/or high-performance distributed storage including Network Appliance, Yahoo!, IBM Almaden Research Center, Symantec, Google, VMware, and others.
- Invited talks—Faculty and students have given a large number of invited talks on this research projects at industry venues and at other universities. This has resulted in an extraordinary amount of interest and direct transfer of some algorithms to Yahoo! and others.

- Collaboration—The members of the research team have active collaborations with representatives from many companies working in the storage industry including IBM, HP, Symantec, and Network Appliance leading to direct transfer of some technologies developed in this project.
- Ceph open source—We have made the Ceph source code publicly available on Sourceforge for download, investigation, modification, and use by anyone. To date the Ceph Sourceforge site has had approximately 40,000 views including thousands per month since two months before the OSDI publication in Dec. 2006.

9 Spin-Off Research

UCSC/LANL Institute for Scalable Scientific Data Management (ISSDM)

The UCSC/LANL Institute for Scalable Scientific Data Management (ISSDM) is a $1M per year collaborative research and education program between UCSC and Los Alamos National Laboratory (LANL). Led by Computer Science Professors Scott Brandt at UCSC and the LANL High Performance Computing Systems Integration Group Leader, Department of Energy/National Nuclear Security Agency I/O and Storage Leader, and High End Computing Inter-agency Working Group (HECIWG) File Systems and I/O Leader Gary Grider and Carolyn Connor-Davenport at LANL, the ISSDM is part of the LANL National Security Education Center (NSEC), whose goals are to recruit, retain, and revitalize Laboratory staff by:

- Developing long-term collaborative relationships with universities whose research interests are important to the Laboratory.
- Sponsoring, partnering with, and funding university professors and students in areas that are important to meet Laboratory objectives.
- Establishing relationships with students working in these research areas and recruiting them to the Laboratory upon graduation where they can continue their work and help the Laboratory to fulfill its objectives.
- Instituting educational programs to provide Laboratory personnel with specific knowledge and skills that make them more effective in completing projects that meet Laboratory objectives.

The ISSDM focuses on Computer Science and Engineering, especially as it relates to scalable scientific data management. The ISSDM supports the NSEC goals through parallel efforts in education and research.

Education—A remote education program allowing degree-seeking and continuing-education students at LANL to take UCSC graduate courses in Computer Science and Engineering. Selected courses are broadcast to LANL via Polycom with 3-6 remote students in each class. The ISSDM pays 15% of the instructor salary and provides the necessary equipment (when not already in Polycom-equipped classrooms) and associated teaching and technical support.

Research— A synergist research program whereby funding is provided to UCSC faculty and students via an internal competitive proposals process. Focusing on areas of mutual interest within UCSC and LANL, the ISSDM will eventually provide up to $600K annually in sponsored research funding to SOE faculty across most or all of the SOE departments. Goals of this funding include facilitating collaborative research in areas of mutual interest, communicating LANL research needs and problems to UCSC and transferring UCSC research and technology to LANL, seeding collaborative projects that will lead to significant external funding, and rewarding faculty for participation in the educational program and supporting the NSEC goals. Current areas include storage systems, database systems, scientific visualization, networking, and machine learning.

Petascale Data Storage Institute (PDSI)

The Petascale Data Storage Institute is a $11.5M collaboration between CMU, Michigan, UCSC, and five national labs funded by the DOE Office of Science.

Petascale computing infrastructures for scientific discovery make petascale demands on information storage capacity, performance, concurrency, reliability, availability, and manageability. The last decade has shown that parallel file systems can barely keep pace with high performance computing along these dimensions; this poses a critical challenge when petascale requirements are considered. The Petascale Data Storage Institute will focus on the data storage problems found in petascale scientific computing environments, with special attention to community issues such as interoperability, community buy-in, and shared tools. Leveraging experience in applications and diverse file and storage systems expertise of its members, the institute

allows a group of researchers to collaborate extensively on developing requirements, standards, algorithms, and development and performance tools. Mechanisms for petascale storage and results will be made available to the petascale computing community. The institute will hold periodic workshops and develop educational materials on petascale data storage for science.

The Petascale Data Storage Institute is a collaboration between researchers at Carnegie Mellon University, National Energy Research Scientific Computing Center, Pacific Northwest National Laboratory, Oak Ridge National Laboratory, Sandia National Laboratory, Los Alamos National Laboratory, University of Michigan, and the University of California at Santa Cruz.

End-to-end Performance Management for Large, Distributed Storage

This project is a $1M NSF-funded research project funded under the HECURA program.

Storage systems for large and distributed clusters of compute servers are themselves large and distributed. Their complexity and scale makes it hard to manage these systems, and in particular they make it hard to ensure that appli- cations using them get good, predictable performance. At the same time, shared access to the system from multiple applications, users, and competition from internal system activities leads to a need for predictable performance. Intellectual merit. This project investigates mechanisms for improving storage system performance in large distributed storage systems through mechanisms that integrate the performance aspects of the path that I/O operations take through the system, from the application interface on the compute server, through the network, to the storage servers. We focus on five parts of the I/O path in a distributed storage system: I/O scheduling at the storage server, storage server cache management, client-to-server network flow control, client-to-server connection management, and client cache management.

Much of the existing work on QoS for storage considers management of the individual elements of the path that I/Os take through a storage system, but little of the work considers end-to-end management of the whole path. The problem with a naive chaining of multiple management algorithms along the path (e.g., one algorithm for the network, and another for the the storage server) is that emergent behaviors that arise from such chaining can reduce the overall performance of the system. Also, much of the existing work is specific to continuous media and other applications with periodic real-time I/O workloads, as opposed to applications with general workloads. The unifying idea in this project is that the storage server should control data movement between clients and the server. Only storage server has knowledge of the I/O demands across all its clients. The server is also more likely to contain a bottleneck resource than any individual client is. Accordingly, the server can make I/O scheduling decisions to balance client usage, can manage cache space taking into account the workload from all clients contending for the cache, and can manage the network flow.

The techniques build on our current I/O scheduling work, which allow applications to specify quality of service for I/O sessions. The QoS includes a reserved (minimum) performance, a limit on performance, and fair sharing of extra performance among sessions. The project extends the scheduling work to improve disk utilization, and then uses QoS and utilization information to guide cache management and network flow control decisions. We also investigate how we can use machine learning techniques to predict near-future resource demand in order to handle those clients that are connected over long-latency links.

These techniques should help storage systems to scale to support the compute clusters currently being planned. Large scale means sharing, both within one application and between applications. Performance management ensures that each application or client gets good performance. For example, when many nodes are computing simulation data and other nodes are visualizing that data, the two can proceed without interference. Large scale also means there will always be system maintenance going on to handle failure and replacement. Performance management ensures that maintenance can proceed without interfering with applications.

File System Tracing, Replaying, Profiling, and Analysis on HEC Systems

This project is a $750K NSF-funded research project funded under the HECURA program.

The PIs propose to develop scalable tools and techniques that will work on large clusters to identify I/O performance bottlenecks, visualize them for cluster users for ease of analysis. Thiese techniques will conduct large scale tracing and replaying, collecting vital information useful to analyze the cluster's performance given a specific application. The PIs will use automated and user driven feedback to raise or lower the level of tracing on individual cluster nodes to (1) zoom in/ on hotspots and (2) trade off information accuracy vs. overheads. Extensive investigations will go toward ensuring the accuracy of the information.

New and small projects

- In-flight data (Prof. Scott Brandt, Dr. Carlos Maltzahn, Anna Povzner and Tim Kaldewey)

- Reliability (Prof. Scott Brandt, Dr. Carlos Maltzahn, David Bigelow)
- Active object storage (Prof. Scott Brandt, Dr. Carlos Maltzahn, Esteban Molina-Estolano)
- Distributed document management (Prof. Scott Brandt, Ian Pye)
- Distributed search (Prof. Ethan Miller, Dr. Carlos Maltzahn, Prof. Yi Zhang)

10 Conclusion

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 ASCI-sponsored research in addressing these problems has resulted in the most successful research project in the history of the UCSC Computer Systems Research Group. It has resulted in 35+ publications, 10+ Ph.D.s and M.S.s, numerous technical accomplishments including a variety of effective data and metadata management solutions and a complete working prototype of Ceph, petascale distributed high-performance data storage system, which has been released to the public under an open-source license.

Driven by the data rates and usage scenarios that characterize the mission of the Defense Programs (DP) Laboratories, our research has successfully addressed a number of the key research issues in the design of a high-performance multi-petabyte storage system targeted for use in post-Purple computing systems. Our research addresses these issues and achieves the overall goal of developing 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 and other general-purpose access patterns. Our results have been validated and demonstrated by the creation of proof-of-concept implementations of this architecture culminating with the development and open-source release of Ceph. A major value of this research is its use by the DP Laboratories in providing direction to the vendors of post-Purple systems in building new file and storage systems.