PETASCALE DATA STORAGE INSTITUTE

Final Report

Garth Gibson, Principal Investigator

Carnegie Mellon University, School of Computer Science

5000 Forbes Ave., Pittsburgh, PA 15213

412-268-5890 (-4740 Jennifer Landefeld, assistant), garth@cs.cmu.edu

DOE/Office of Science Program Office: Advanced Scientific Computing Research (ASCR)

DOE/Office of Science Program Office Technical Contact: Lucy Nowell

DOE Grant Number: ***DEFC0206ER25767 (CMU)***

Collaborating Institutions/PI:

- Carnegie Mellon University (Institute lead), Garth Gibson (Institute PI)
- University of California at Santa Cruz, Darrell Long
- University of Michigan at Ann Arbor, Peter Honeyman
- Los Alamos National Laboratory, Gary Grider
- National Energy Research Scientific Computing Center, William Kramer, John Shalf
- Oak Ridge National Laboratory, Philip Roth
- Pacific Northwest National Laboratory, Evan Felix
- Sandia National Laboratory, Lee Ward

Abstract:

Petascale computing infrastructures for scientific discovery make petascale demands on information storage capacity, performance, concurrency, reliability, availability, and manageability. The Petascale Data Storage Institute focuses 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.

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.

Table of Contents

1 Introduction

1.1 Post-PDSI Highlights

In retrospect, the Petascale Data Storage Institute's most innovative and impactful contribution is the Parallel Log-structured File System (PLFS). Published in SC09, PLFS is middleware that operates in MPI-IO or embedded in FUSE for non-MPI applications. Its function is to decouple concurrently written files into a per-process log file, whose impact (the contents of the single file that the parallel application was concurrently writing) is determined on later reading, rather than during its writing. PLFS is transparent to the parallel application, offering a POSIX or MPI-IO interface, and it shows an order of magnitude speedup to the Chombo benchmark and two orders of magnitude to the FLASH benchmark. Moreover, LANL production applications see speedups of 5X to 28X, so PLFS has been put into production at LANL. Originally conceived and prototyped in a PDSI collaboration between LANL and CMU, it has grown to engage many other PDSI institutes, international partners like AWE, and has a large team at EMC supporting and enhancing it. PLFS is open sourced with a BSD license on sourceforge. Post PDSI funding comes from NNSA and industry sources. Moreover, PLFS has spin out half a dozen or more papers, partnered on research with multiple schools and vendors, and has projects to transparently 1) distribute metadata over independent metadata servers, 2) exploit drastically non-POSIX Hadoop storage for HPC POSIX applications, 3) compress checkpoints on the fly, 4) batch delayed writes for write speed, 5) compress read-back indexes and parallelize their redistribution, 6) double-buffer writes in NAND Flash storage to decouple host blocking during checkpoint from disk write time in the storage system, 7) pack small files into a smaller number of bigger containers.

There are two large scale open source Linux software projects that PDSI significantly incubated, though neither were initated in PDSI. These are 1) Ceph, a UCSC parallel object storage research project that has continued to be a vehicle for research, and has become a released part of Linux, and 2) Parallel NFS (pNFS) a portion of the IETF's NFSv4.1 that brings the core data parallelism found in Lustre, PanFS, PVFS, and Ceph to the industry standard NFS, with released code in Linux 3.0, and its vendor offerings, with products from NetApp, EMC, BlueArc and RedHat. Both are fundamentally supported and advanced by vendor companies now, but were critcally transferred from research demonstration to viable product with funding from PDSI, in part. At this point Lustre remains the primary path to scalable IO in Exascale systems, but both Ceph and pNFS are viable alternatives with different fundamental advantages.

Finally, research community building was a big success for PDSI. Through the HECFSIO workshops and HECURA project with NSF PDSI stimulated and helped to steer leveraged funding of over $25M. Through the Petascale (now Parallel) Data Storage Workshop series, www.pdsw.org, colocated with SCxy each year, PDSI created and incubated five offerings of this high-attendance workshop. The workshop has gone on without PDSI support with two more highly successfully workshops, rewriting its organizational structure to be community managed. More than 70 peer reviewed papers have been presented at PDSW workshops.

1.2 Executive Summary

The Petascale Data Storage Institute brings together high performance file and storage system expertise and experience meeting the high performance storage requirements of today's DOE terascale scientific discovery through advanced computing for the purpose of identifying, resolving and setting in motion solutions for the storage capacity, performance, concurrency, reliability, availability and manageability problems arising from petascale computing infrastructures for scientific discovery.

Led by Carnegie Mellon University, the Petascale Data Storage Institute membership also includes University of California at Santa Cruz, University of Michigan at Ann Arbor, Los Alamos National Laboratory, National Energy Research Scientific Computing Center, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, and Sandia National Laboratory.

In this report, the Institute's work will be organized into three functions:

1. Dissemination (Outreach and Standards)
2. Data Collection (Performance Characterization and Failure Characterization)
3. Exploration

Because the Institute focuses on low level files systems and storage systems, its role in improving SciDAC systems was one of supporting application middleware such as data management and system-level performance tuning.

The Institute's most significant contributions at the time of its completion were:

- Dissemination: Extensive delivery of SciDAC system understanding and requirements to academic and commercial file system and storage developers through petascale storage at least two workshops per year, 20-30 different students trained per year, and facilitation of $25M+ additional funding targeted at SciDAC storage systems by NSF and DOD programs.
- Dissemination: Contributions to and guidance of the standardization of parallel file system client software and communications protocol in the IETF's NFSv4.1 Parallel NFS. pNFS is being developed by a large group of commercial vendors based on our longterm leadership. IETF draft standard has been achieved; Linux 2.6.30 has adopted the first of the implementation code with the intention to adopt the rest as it is finished; and RedHat is normalizing its kernel interfaces for supporting pNFS. The Institute is now providing SciDAC-scale testing of pNFS implementations to guide future commercial offerings to met SciDAC needs.
- Data Collection: Extensive failure data collection and publication, up to a decade of terascale failure data from NERSC, LANL, Sandia, PNNL, PSC, Cray, and others coming, and failure data analysis spurring renewed attention to weaknesses in SciDAC system fault tolerance strategies, specifically checkpoint capture bandwidth requirements. This data includes storage failures but is much broader because storage is a core part of extreme-scale fault tolerance for all failure types.
- Exploration: A new checkpoint-optimized stacked parallel file system, Parallel Log-structured File System (PLFS), that transparently converts highly concurrent, small, strided N client to 1 file write patterns that can be totally non-scalable on many of SciDAC's deployed parallel file systems into encapsulated sequential streaming N client to N log file write patterns that scale on all of SciDAC's deployed parallel file systems. Testing at large scale has shown promising results: a PLFS solution may deliver 10X checkpoint capture speeds while often requiring no changes in SciDAC applications

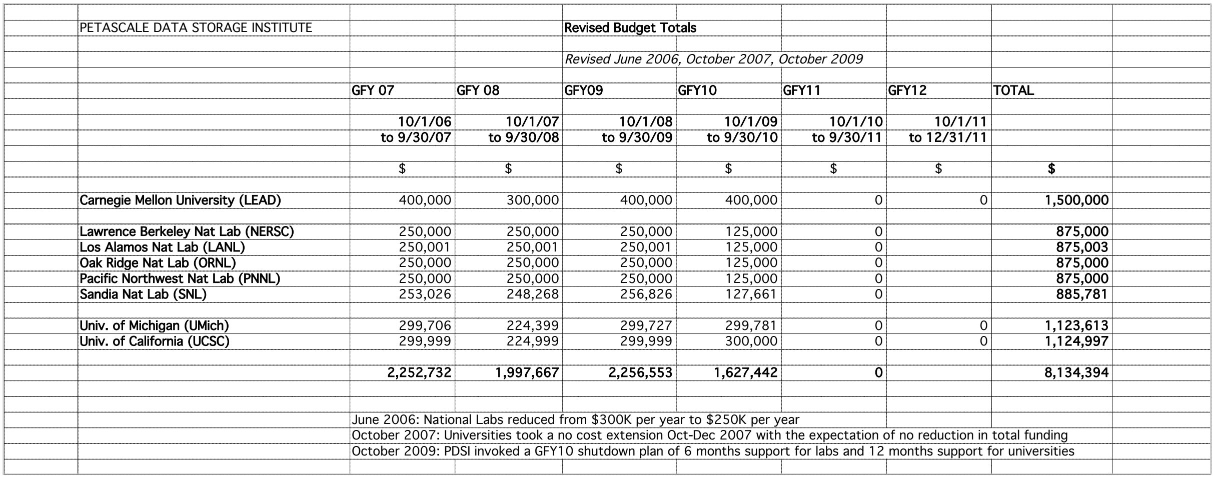
These and other contributions will be further explained below. The total supported personnel, publications and outreach is summarized in the following table.

	FY07	FY08	FY09	FY10	FY11	Total
Faculty/Staff	27	28	26	23		104
Students	28	32	33	22		115
Total Supported	55	60	59	45	0	219
Journal	5	4	5	1		15
Conf+Worksp	27	27	26	16	4	100
Other pubs	6	49	22	5	3	85
Total pubs	38	80	53	22	7	200
Talks	99	69	61	29		258
Workshops	1	2	2	2	1	8

1.3 Budget Summary

The Institute originally requested $12.5M over a five-year program. The eight collaborating universities and laboratories requested approximately $300,000 annually, with the exception of the lead institution, Carnegie Mellon University, which requested approximately $400,000 annually to accommodate its larger leadership coordination role in addition to its full research and outreach roles. When the Institute was funded, the five national laboratories were reduced by about $50,000 per year. In the beginning of GFY 2008 the three universities agreed to take a no cost extension through December 2007, shifting 25% of GFY08 budget out to GFY12. In September 2009 ASCR requested that PDSI revise its plans to complete in FY2010 with each lab spending 50% of current budget and each university spending 100% of current budget.

The following table is the resulting funding profile.



PETASCALE DATA STORAGE INSTITUTE			Revised Budget Totals				
			Revised June 2006, October 2007, October 2009				
	GFY 07	GFY 08	GFY09	GFY10	GFY11	GFY12	TOTAL
	10/1/06 to 9/30/07	10/1/07 to 9/30/08	10/1/08 to 9/30/09	10/1/09 to 9/30/10	10/1/10 to 9/30/11	10/1/11 to 12/31/11	
	$	$	$	$	$	$	**$**
Carnegie Mellon University (LEAD)	400,000	300,000	400,000	400,000	0	0	**1,500,000**
Lawrence Berkeley Nat Lab (NERSC)	250,000	250,000	250,000	125,000	0		**875,000**
Los Alamos Nat Lab (LANL)	250,001	250,001	250,001	125,000	0		**875,003**
Oak Ridge Nat Lab (ORNL)	250,000	250,000	250,000	125,000	0		**875,000**
Pacific Northwest Nat Lab (PNNL)	250,000	250,000	250,000	125,000	0		**875,000**
Sandia Nat Lab (SNL)	253,026	248,268	256,826	127,661	0		**885,781**
Univ. of Michigan (UMich)	299,706	224,399	299,727	299,781	0	0	**1,123,613**
Univ. of California (UCSC)	299,999	224,999	299,999	300,000	0	0	**1,124,997**
	2,252,732	**1,997,667**	**2,256,553**	**1,627,442**	**0**		**8,134,394**

June 2006: National Labs reduced from $300K per year to $250K per year
October 2007: Universities took a no cost extension Oct-Dec 2007 with the expectation of no reduction in total funding
October 2009: PDSI invoked a GFY10 shutdown plan of 6 months support for labs and 12 months support for universities

1.4 Management Plan

The Petascale Data Storage Institute was a university-led distributed center involving multiple institutions, including universities and DOE National Laboratories. It was led by Carnegie Mellon University with Garth Gibson as its Principal Investigator with overall coordination responsibility. The Institute's work was organized into six projects of three types, each initially with its own project leader. This organization was revised around the natural leadership of an organization's co-principal investigator, with direct oversight by the Institute PI. Conference calls were held approximately once a month, and face-to-face meetings were collocated at events most members attended. Specifically, there were at least three annual face-to-face meetings collocated with SCxy, FAST, and FSIO conferences and workshops. Additionally a mangers' mailing list provided the bulk of the coordination.

Source code for software deliverables is fully and freely available for use and modification throughout the scientific computing community unless otherwise declared. The open source license used varies across deliverables, but is one of the licenses approved and certified by the Open Source Initiative (http://www.opensource.org), such as the classic GPL, LGPL, BSD and MIT licenses.

Our strategy for long term support for many of our deliverables is to achieve adoption by the Linux community. While this cannot be guaranteed a priori, it is working for Parallel NFS. Other deliverables are open source projects distributed as is and are available to the HPC community at large. PLFS, in particular, (http://sourceforge.net/projects/plfs) is supported by a team including Los Alamos and EMC and is open source under a BSD license.

1.5 Background for Merit Review

As suggested in the SciDAC program notice, this section identifies where and how each merit review criterion is addressed.

1.5.1 Scientific and/or Technical Merit of the Project.

a) *Potential for significant impact on SciDAC applications*. Storage performance is central to the effective use of extreme scale machines at least because of checkpoint fault tolerance. A recent analysis showed S3D spent 1% of runtime in I/O at 512 cores, but 30% at 16,000 cores. This is a trend that can be endured as we scale to millions of cores.

b) *Demonstrated capabilities of the applicants*. The Institute reflects an exceptionally strong record of basic research, its transition to practice and operations of SciDAC class systems. Our team includes access and management roles in key DOE petascale computing facilities (NERSC, PNNL, LANL, SNL and ORNL). Its academics and partners have shaped most of the parallel file systems deployed for SciDAC use (Lustre, PanFS, PVFS). Additionally, Institute members have made essential contributions to the open source community, especially to NFS and Parallel NFS implementations.

c) *Coupling with scientific simulation*. Institute membership spans National Labs staff who directly support massive scientific computations as well as researchers who work with scientific simulation themselves on a regular basis.

d) *Impact on science disciplines outside of SciDAC applications*. The central role of storage in extreme scale fault tolerance is much broader than SciDAC applications. Maintaining balanced systems while top500.org systems' speed is growing faster than transistor density, and storage device speeds are growing much slower, is a problem that cannot be solved once; it manifests a challenge with each new generation of systems.

e) *Approach to long-term support and transfer*. We have a strong record of moving our code into community supported open source, specifically Linux and RedHat; and into commercially supported systems.

f) *Broad community interaction*. Institute members span the state of the art in storage systems research and practice, featuring top National Lab staff and the leading academic storage systems research centers. All of the Institute members have enjoyed long and productive interactions with industry partners; the academic members are collaborating directly with (and supported in part by) a broad collection of the storage industry, including IBM, HP, Intel, Sun, Panasas, EMC, Microsoft, Network Appliance, Google, Yahoo, Facebook, Veritas, Seagate and others. Additionally, the team enjoys close interaction with supercomputer and file system industry partners. The PDSI was a leader in community building in the file systems and I/O area for high end computing.

1.5.2 Appropriateness of the Proposed Method or Approach.

a) *Plan for coupling to emerging advances in enabling technology or to applications researchers*. Our strong outreach, though academic (FAST), HEC (SCxy) and government (FSIO) workshops gave us broad forums for coupling. New data repositories, performance tracing tools and transparent checkpoint systems (PLFS) facilitated interactions with applications researchers. Existing coupling between Institute members and the storage R&D community also facilitates these connections, especially for Parallel NFS.

b) *Approach to intellectual property management and open source licensing*. Dissemination and data collection was done fully in the open in almost all cases. Reference implementations, tools, and traces produced as part of the Institute's work are being shared via open source, as discussed in the Management Plan.

c) *Plan for effective collaboration among participants*. Frequent face-to-face meetings, and a variety of collaborative work items, specifically in data collection and outreach, facilitated collaboration. The need to understand SciDAC applications and test novel mechanisms at scale drew academics to seek out collaborations with lab members.

d) *Plan for ensuring communication with other efforts*. The outreach and dissemination effort of the PDSI was explicit about ensuring communication with other stakeholders exploring and supporting massive-scale science applications. Liaison's with key partners: Rob Ross with SDM and Phil Roth with PERI, for example, played an important role.

1.5.3 Competency of Applicant's Personnel and Adequacy of Proposed Resources.

The Institute brought together leading researchers from the top academic storage systems research centers and lead staff managing large-scale storage at National Labs.

As to resources, some of our efforts depended on external funding, such as Parallel NFS, or the collection and publishing of failure data by non-Institute HEC sites. We had good success with finding these partnerships. Some unplanned activities were perhaps short of resources – continued availability of testing-at-scale facilities was always a challenge. Perhaps the biggest opportunity missed by the early shutdown of PDSI would have been additional funds for hardening and packaging PLFS, and transitioning it to an open source maintainer and commercial supporter. Luckily other DOE and non-DOE programs have stepped up and advanced PLFS. Additionally, some additional funds for better project management would have been of great value to the PDSI members and especially to the SciDAC program.

1.6 Close Out Plan

On September 2, 2009, Daniel Hitchcock informed us that PDSI was to come to an orderly conclusion over the next 12 months..

Close Out Plan for Petascale Data Storage Institute (PDSI)

PDSI was brought to an orderly conclusion over FY2010, with 6 months FY2010 support to PDSI laboratory partners and 12 months FY2010 support to PDSI university partners. All staff supported by SciDac were transitioned to other funding sources or terminated as determined by each participant institution. As the laboratories received a partial year allocation, the duration of their close out varied based on their planning. PNNL and LANL stretched the allocation over all of FY2010, while SNL and ORNL continued at the normal burn rate and completed with 6 months of effort. NERSC had an intermediate plan completing at the end of May 2010.

Ongoing outreach activities will no longer be supported by SciDac, but did not expected to cease. In particular, the SCxy-collocated Petascale Data Storage Workshop has been growing in attendance, and is typically one of the best-attended workshops at SCxy. It has transitioned to become a community led activity. Also, the HECIWG/FSIO workshop held in DC in August that has been used for an annual NITRD commissioned research roadmap review will no longer be supported by SciDac PDSI. While it did continue under NSF leadership for awhile it has since terminated. Continued penetration of HPC Science considerations into academic curricula will cease as a SciDac PDSI funded activity, but is expected to continue based on its independent merit.

Code artifacts deemed complete by contributing institutions were be released into open source where depending licenses allow and linked into PDSI web materials. Some tools, notably Pergamum, have been picked up by start up companies. A few incomplete code projects have been identified as compelling to transition to other support. Specifically, the Parallel Log-structured File System (PLFS) which is delivering one and two orders of magnitude speed up of production science checkpointing, continued to be developed and hardened for deployment, primarily with NNSA support and a partnership with EMC. Also, the development and shepherding of Parallel NFS (pNFS) implementation in Linux is of importance to users and markets beyond SciDac applicationss, and continued under support from companies such as EMC, NetApp, Panasas and IBM and agencies such as the DoD. Other partially developed activities, especially at universities, continued until a natural completion or handoff point with internal funding sources, as deemed by the developer institutions.

More specifically, PNNL continued repeating activities such as data collections and outreach, and code support and documentation through FY2010, and closed out more promptly work on IOMMU and VMs embedded in hardware RAID systems. UCSC transitioned, in part, its metadata related research into the PVFS (SciDac SDM artifact), and its storage security research results into PVFS or the open source Ceph project. SNL transitioned its S3D over HDF5 accelerations to a funded collaborator at Nortwestern University, terminated its less advanced efforts on CCSM performance, and put most of its effort into completing the Compute Node Linux low-overhead IO tracing tool. LANL contributed to the most impactful outreach activities in FY2010, updated data collections, and hardened and transitioned PLFS, already an open source project on sourceforge, for wider use. ORNL closed out with two research reports on scalable tracing and an application of its performance prediction framework, including IO, to two more SciDac applications. CMU led outreach and data collection activities as long as others are contributing, continued its integration of the PLFS platform into an academic research vehicle for wider stimulation of related HEC research, and found an alternative funding source for scalable metadata research. CMU shutdown or transitioned to other sources predictable performance work and failure diagnosis and recovery work to enable more progress on PLFS and scalable metadata. NERSC terminated or transitioned the tape reanalysis project and completed the PDSI/PERI collaboration on IO profiling project and the FLASH storage in HPC systems project. University of Michigan continued its shepherding of Parallel NFS into the Linux kernel and and transitioned funding to industry sources.. HPC scalability testing continued during FY2010, but the performance vs reliability tradeoffs project and exploitation of modern chip sets for IO project were terminated.

2 Dissemination

2.1 Outreach

The SciDAC2 solicitation for University led Institutes required that SciDAC Institutes be partially outward focused towards R&D communities and provide much outreach, to help connect SciDAC to a broader High Performance Computing (HPC) R&D community. Of course to perform this communications/outreach function, it was vital to have the SciDAC program be grounded in the SciDAC site and application needs and to have many different high quality outreach mechanisms. The Institute took this outreach request quite seriously and the following section of this document outlines many of these outreach mechanisms and activities.

2.1.1 Guiding the National Research Agenda and National Coordination of R&D Investments

The PDSI provided funding to manage the High End Computing File Systems and I/O (HECFSIO) national coordination effort. The HECFSIO is a technical advisory group which reports to the HEC Interagency Working Group (HECIWG), a multi-agency committee formed to coordinate government agency investments in HEC R&D. To help plan for the research needs in the area of File Systems and I/O, the HECIWG designated FSIO as an area of national focus starting in FY06. To collect a broader set of research needs in this area, the first HECFSIO workshop was held in August 2005 in Grapevine, TX. Government agencies, top universities in the I/O area, and commercial entities that fund file systems and I/O research were invited to help the HECIWG determine the most needed research topics within this area. The workshop attendees helped catalog government funded FSIO R&D and prioritize gaps in need of R&D funding. Since 2005, the HECFSIO group has helped coordinate government funded R&D including coordination of:

- 2006 $12M+ NSF HECURA FSIO solicitation which resulted in 22 university collaborative research projects in the gap areas
- 2007 $1.5M NSF CPA - 5 FSIO Projects
- 2007 $1M DOD ACS - 1 FSIO projects
- 2008 $0.5M NSF CPA - 2 FSIO projects
- 2009 $10M+ NSF HECURA FSIO solicitation which will result in many collaborative research projects in the remaining gap areas. The solicitation will concentrate on areas that are still considered to be gaps and will seek to round out the overall portfolio of R&D to cover the gaps as well as seek proposals which are both evolutionary and revolutionary.

Additionally, the HECFSIO coordination effort listens to what the R&D community needs beyond funding, like access to operational data, traces, and workload information as well as access to test beds and HEC sites for collaboration.

The heart of the HECFSIO coordination effort revolved around an annual PDSI sponsored workshop for HECFSIO which brings together government HEC sites and University and Industry HECFSIO researchers to help keep the R&D portfolio balanced and well targeted at the most pressing R&D needs. To provide needed documentation of the HECFSIO needed R&D as well as a listing of existing HECFSIO coordinated projects, the HECFSIO coordinators maintain an annual document release with the HECFSIO R&D Roadmap which is used to convey the R&D portfolio and the most pressing needs. The HECFSIO information can be found at http://institute.lanl.gov/hec-fsio. Without the PDSI, the HECFSIO effort could have been severely impacted in its ability to provide coordination of HECFSIO R&D for the nation. DOE NNSA and the NSF thankfully stepped in and provided funding to keep this activity alive in the termination of PDSI funding. After a few more meetings, however, the HECFSIO coordination has ceased.

2.1.2 Workshops and Tutorials

A big part of the PDSI strategy for outreach is information dissemination. PDSI chose to attack this mission in many ways, the most notable appear below.

Workshops and Conferences PDSI Organized

The premier HPC conference is of course Supercomputing, and the PDSI has taken advantage of Supercomputing by providing fastest growing in attendance workshop in the history of Supercomputing, the Petascale Data Storage Workshop (PDSW) has been provided by the PDSI at SC06, SC07, SC08, and

SC09. Each year the workshop was standing room only with more attendance each year. These workshops have gotten the word out about HPC storage related activities provided the opportunity for dozens of HPC storage related talks, posters, and short papers, and panels as well as provided a community building forum. Also at Supercomputing, multiple Birds-of-a-Feather (BoF) and panels have been sponsored by PDSI including multiple NFSv4/pNFS related BoFs and a standing room only "Exa and Yotta Scale Data - Are We Ready?" panel at SC08, the last panel of the conference on Friday, with every seat filled. Suffice it to say, the PDSI Supercomputing footprint was significant, successful, and growing. The members of the PDSI intend for some of this work to continue even without PDSI/SciDAC support, via funding from other sources.

Other workshop like activities of note include:

- Multiple BoFs at the USENIX File Systems Conference (FAST) on various topics like HECFSIO and NFSv4/pNFS.
- An I/O benchmarking excellence workshop held at UCSC
- Multiple NFSv4 interoperability workshops held by CITI
- NFSv4/pNFS tutorial at USENIX 2007

Of course, as mentioned above, each year the PDSI sponsors the HECFSIO conference. This is the premier HPC storage and I/O government funded R&D event held each year. Also, with members of the SDM Center, ORNL PDSI team members submitted an HPC I/O tutorial to the 2008 ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP) which was accepted.

2.1.3 Education

Another facet of the PDSI's approach to outreach is in the area of education. Education of the next generation of HPC experts is vital and should be a part of the job of every current HPC expert. The PDSI worked in several areas outlined below to outreach HPC related graduate education.

Enhancing Curriculum and Participation

The PDSI enhanced education in Computer Science to contain more HPC storage related material. Two methods were explored to have an effect on education, curriculum enhancement and HPC site experts spending time lecturing in college classes.

Multiple successful curriculum modifications have been accomplished via PDSI activities. CMU developed HPC material for graduate course on Storage Systems. Follow on research was done by multiple graduate students. The second offering of this class material was planned with assistance from SDM Center researchers. The third version of this class material was put together based on PLFS.

Additionally, the PDSI had a strong ally in the educational outreach area. The LANL/UCSC Institute for Scalable Scientific Data Management (ISSDM), an educational institute between LANL and UCSC, with strong assistance from PDSI, helped enable additions to the UCSC graduate storage class with HPC content. The ISSDM provided distance learning capabilities to allow delivery of HPC content from HPC sites into the UCSC classroom.

Additionally, the PDSI helped initiate a new parallel programming class at the Colorado School of Mines which has parallel I/O content provided by the PDSI. Also, with help of the ISSDM and the PDSI, the University of New Mexico is planning on including parallel storage in some of its systems programming courses.

Lecturing in classes to deliver HPC storage content and experiences is also an important part of the PDSI education outreach activities. Many lectures have been provided to universities to enhance college courses with HPC storage related information by PDSI members. UC Berkeley, New Mexico Tech, Colorado School of Mines, Amherst, UCSC, and CMU have all received the benefit of this activity.

Supporting Faculty and Students

The numbers of PDSI supported personnel is listed in Metrics section and names are given in the appendix. In each year a total of 23-28 faculty or research staff were supported and a total of 45-60 graduate and undergraduate students were supported.

Another PDSI educational outreach was to provide additional funding for extra-PDSI students and faculty at universities to work on HPC storage related projects. Recipients included personnel at UC Berkeley, UCSC, CMU, Michigan, Colorado School of Mines, New Mexico Tech, and others. In 2008, 8 faculty and 28 students were funded some amount, and in 2007, 13 faculty and 30 students were funded some amount. Additionally, the PDSI provided support for 5 students to attend USENIX FAST07, FAST08, and FAST09.

Additionally, via the ISSDM and the LANL/CMU Institute for Reliable High Performance Information Technology (IRPHIT), over a dozen students were mentored via distance learning/video conferencing capabilities. This partnership between PDSI and ISSDM/IRHPIT helped to get the HPC storage word out into higher education.

Equipment

In order for universities to do HPC oriented research, it is vital that universities have access to HPC type computing, networking, and storage resources. The PDSI assisted the ISSDM and IRHPIT to provide some HPC related network and storage equipment including Infiniband, Ethernet, compute, and storage test equipment to assist in HPC related research and education.

2.1.4 Support for other SciDAC Activities

As was stated above, one important part of the PDSI outreach program was to reach in to other parts of SciDAC to disseminate PDSI work, collect input, and collaborate with applications, sites, and other SciDAC researchers/Institutes/Centers.

SciDAC Workshops and PI Meetings

An important vehicle the SciDAC program provides to disseminate information and promote collaboration is the SciDAC PI workshop and meetings. PDSI has participated fully in these useful and important SciDAC events. PDSI thrust areas like failure data release and research, metadata research, and I/O tracing and analysis of SciDAC applications have been featured at these workshops and meetings.

2.1.5 Collaboration with Institutes and Centers

Since the PDSI focus was file systems and storage, which resides in the HPC software stack below applications and data management layers, a subset of the Scidac Institutes and Centers are candidates for collaboration. PDSI's collaborations have been primarily with the Performance Engineering Research Institute (PERI) and Scientific Data Management Center for Enabling Technologies (SDM).

<u>PERI Performance Engineering Research Institute</u>

Given that PERI's goal was to work with SciDAC applications on performance for the most part, the collaborations between PDSI and PERI have concentrated on SciDAC application performance issues. PDSI primarily was interested in I/O and storage related performance characterization and assistance. PERI was interested in other non I/O and storage related performance. PDSI's work in characterizing I/O of SciDAC applications has been shared with PERI as an ongoing outreach activity to ensure a coordinated performance effort. Performance and characterization information on the S3D application on the ORNL Cray XT system was one of the prime focuses of this collaboration. Other applications such as POP have been characterized as well, so more information sharing with the PERI team on application I/O characterization and performance modeling will be forth coming. The PDSI team also integrated an ORNL developed I/O measurement and analysis infrastructure into a performance prediction framework that is being developing by the PERI team.

SDM Scientific Data Management Center

Given that the SDM Center was focused on data management software, much of which utilizes file systems and I/O software that the PDSI was most interested in, it makes perfect sense that the SDM Center and PDSI were collaborators. There have been several notable collaborations between these two entities. Bi-directional exchanges in the areas of I/O visualization tools, tracing tools, and I/O kernels from applications of interest occured frequently. Explorations in extreme metadata studies, billions of files in a single file system directory, have been done using the PVFS file system on conjunction with the SDM Center researchers. Standards explorations in the areas of pNFS and POSIX HEC extensions have been done jointly between PDSI and SDM Center researchers. PDSI researchers also worked with the SDM project in relation to Active Storage and its uses in SDM solutions. Additionally a number of technical exchanges have occurred from seminars given by SDM Center researchers at PDSI venues and PDSI researchers attending SDM Center meetings. This was a strong and important collaboration for these two entities.

2.1.6 Collaborations with Sites and Applications

In order to be impactful to the SciDAC mission, it was vital that the PDSI reach out to SciDAC computing sites. PDSI had an excellent relationship NERSC and PNNL storage operations teams and has leveraged these relationships to dialogue about important site operations issues and about PDSI developments. While a lot of effort in SciDAC goes towards working with applications, we felt it wise to also work with the computing site operations to understand their issues as well.

Obviously to be relevant for the SciDAC program, working with applications is vital, both directly and indirectly. Given PDSI's position in the HPC software stack, often much of the application interaction we have is through a software layer that the SDM Center is interested in, which as mentioned above is why PDSI and SDM are close collaborators. Direct interaction with the SciDAC codes is also important though. The following discussion summarizes direct application interactions.

An important first step in working with scientific applications is to get characterization information on how the applications do I/O. A number of interactions with codes and code teams was done. Characterization, tracing, and I/O kernels have been developed or extracted from the Parallel Ocean Program (POP), the Hybrid Coordinate Ocean Model Home Page (HYCOM), the All-Orders Spectral Algorithm (AORSA), the FLASH astrophysics, the S3D turbulent reacting, and the Community Climate System Model (CSSM) codes.

In addition to characterization and publishing of characterization information for these codes, PDSI also developed trace visualization tools to look at the characterizations to help work with applications to improve I/O. The outcome of looking at I/O characterization has resulted in a novel development by the PDSI researchers called the Parallel Log Structured File System (PLFS) which is outlined in the novel section of this document. This tool has demonstrated improvements in unaligned I/O write patterns from real SciDAC applications of over 10 times speed up in I/O. This tool requires no application changes and is less than 3000 lines of code in user space.

2.1.7 File System Research Software Packaging and Distribution

One general support activity is to support other HPC site work with various parallel file systems, PNNL has created and made available for download packages that support PVFS and Lustre on the CentOS, Debian, and Ubuntu Linux distributions. These packages have been downloaded over a thousand times, and the team has also handled various support requests from some of the uses of these packages.

2.1.8 Web of Web Sites

Another outreach mechanism was use of the web for information dissemination. The official PDSI web site http://www.pdsi-scidac.org/ is a top level view of the PDSI with events, dta/code release, and related links. This top level site has pointers to each institution in the PDSI, all of which have PDSI web presence. Additionally, the HECFSIO web site was partially sponsored by the PDSI and partially by the IS-SDM. Some of the most important web presence for PDSI is in the area of making data/tools/codes available for download. There are multiple repositories all linked by the common PDSI top web site that host this technical released data.

2.2 Standardization

One of the primary dissemination categories intended to make long lasting change in the HPC industry and community was the promotion of standards. SciDAC and all HPC applications and systems depend on standards for interoperability and longevity. It is also vital that standards evolve to allow applications and sites to exploit new hardware, software, and solutions.

While we initially foresaw many shallow activities in this area, in fact each activity has significant overhead. The main one, which is having very good success, is Parallel NFS (pNFS). An IETF RFC came out in 2010 and a complete Linux client implementation was released in 2011. We anticipate an outcome of the application of pNFS to SciDAC applications or sites in the near future. Additionally, our activity in our second area of standardization, high end extensions to POSIX application interface, has picked up some steam with some of the proposed standards being implemented in Linux.

NFSv4/pNFS

One of the most prevalent standards in the storage and file systems area is the IETF Network File System (NFS) standard. NFS is a file/data movement/management protocol that has served the computing community well. There is an enormous industry that has been built up around the slowly evolving and quite stable NFS standard. HPC computing sites are dependent on NFS for one of the most important file sharing mechanisms. The evolution of the NFS standard to support a more secure environment is the focus of the NFSv4 effort. DOE has been involved in the NFSv4 work since its inception over half a decade ago. More recently, DOE asked the NFS community to consider scalability in the NFS standards and reference implementation effort. pNFS is an extension to NFSv4 that allows clients to overcome NFS scalability and performance barriers. Like NFS, pNFS is a client/server protocol implemented with secure and reliable remote procedure calls. A pNFS server manages storage metadata and responds to client requests for storage layout information. pNFS departs from conventional NFS by allowing clients to access storage directly and in parallel. By separating data and metadata access, pNFS eliminates the server bottlenecks inherent to NAS access methods.

While the pNFS protocol allows parallel access to files stored in any kind of back end, the IETF working group focuses on access to NFS servers, object storage, and block storage. The generic aspects of the pNFS protocol are described in a working document that is moving toward standardization as the NFSv4.1 specification. That document also describes the pNFS file layout protocol. Mechanisms for access to object and block storage are described in separate documents, which are also moving toward standardization.

By combining parallel I/O with the ubiquitous standard for Internet filing, pNFS promises state of the art performance, massive scalability, and interoperability across standards-compliant application platforms. The University of Michigan Center for Information Technology Integration (CITI), a PDSI member institution, is one of a handful of primary contributors to the Linux-based, open source implementation of NFSv4.1 and pNFS.

PNFS could become a common scalable file system client to our favorite parallel file systems, GPFS, Lustre, PVFS, and Panasas. There are products with these back ends. It is vital that the HPC community follow through with the pNFS effort by testing pNFS implementations at some reasonable scale with HPC workloads and feed back to ensure that the pNFS work will address the needs of the HPC community and eventually SciDAC apps and systems. This has been nearly a decade of effort partially funded by DOE and the most recent support from the PDSI with the PDSI partner CITI was where the big payoff really began, as products based on this technology are nearly ready for release. After PDSI funding was exhausted, DOD provided some support, but the bulk of the support has been provided by vendor companies.

See http://www.pdl.cmu.edu/pNFS/index.html and http://pnfs.com for more information on pNFS.

High End Computing Extensions to POSIX

The POSIX API is "unnatural" for high-end computing applications. Opportunity abounds to make the POSIX I/O API friendlier to HPC, clustering, parallelism, and high concurrency applications. The entire set of operations should be combed over and carefully, consistently, enhanced for high-end computing, clustering, and high concurrency needs, while maintaining complete compatibility for legacy applications.

The PDSI undertook another long term standards effort to try to evolve the POSIX standard. The goal for this effort is to achieve a well accepted by industry POSIX I/O API extension, or set of extensions to make the POSIX I/O API more friendly to HPC, clustering, parallelism, and high concurrency applications. This extension effort will need to be done in phases, due to the fact that some enhancements to the POSIX I/O API to help these applications are well understood, like high concurrence extensions, while other proposed enhancing ideas are not well fleshed out yet, like active storage concepts. Progress has been made on this fledgling standards evolution effort. PDSI in coordination with the SciDAC SDM Center and ANL have performed tests on approximations of various POSIX extensions to demonstrate the performance advantages of several of the extension areas. LANL with contributions from ANL and its PDSI partner institutions has pushed to get PDSI and DOE a seat at the standards table by making the PDSI a member of the Xopen/Open Group which owns the POSIX standard and has written the business justification and provided an initial set of POSIX man pages for the proposed new POSIX functionality. One of the POSIX HEC extensions has been accepted to be included in a future version of the POSIX standard so far. This new function allows applications to query parallel layout information from a file system about a file, allowing applications the ability to optimize I/O patterns. It is expected that this, like almost all standards efforts will take many years to have an effect, but once standardized, new function can be counted on by the HPC community for decades into the future.

See http://www.opengroup.org/platform/hecewg/ for more information on the HEC Extensions for POSIX.

General Standardization Efforts

The increasing use of global and parallel file systems places a great burden on vendors and open source authors to support a diverse and rich collection of client platforms. Most of the tasks accomplished by any such file system are similar, though. An active collaborative research program into guiding efforts to generate common, well-accepted APIs and protocols is required for the health of the HPC storage industry. Additionally, as new concepts in storage and I/O mature beyond prototypes, existing APIs must be enhanced or new APIs need to be developed and validated with real science applications. Such research would strive to support the most important workloads while retaining an ability to be easily extended so that competition via unique features and capabilities is retained.

One of the strengths of this SciDAC Institute was its extremely knowledgeable and influential member institutions and staff. The Institute drew on the science applications experience and diverse file and storage systems experience at the national laboratory partners and the scalable storage research experience of its university partners. Many of these organizations are already collaborating to guide some important standards and API related activities. Additionally, many of the Institute's staff have a track record for successfully influencing industry accepted standards and APIs including technologies like RAID; IETF

NFSv4, pNFS, iSCSI; ANSI/T10 OSD; and others. The establishment of this SciDAC Institute enabled much more extensive collaborative exploration into trade-offs which will provide the basis for guiding the development of standards and APIs for petascale storage. The Institute also enabled reference implementations of HPC storage standards and APIs and facilitate validation of these using real science applications. This powerful combination of knowledge, influence, reference implementations, and meaningful validation gave this Institute the unique capability to move HPC storage related standards and APIs forward within the standards communities and industry.

3 Data Collection

3.1 Summary

We cannot predict the required characteristics of future storage and I/O infrastructure without understanding the storage and I/O demands of current applications. However, obtaining that understanding has proven difficult for large-scale scientific applications due to lack of appropriate benchmarks, activity traces, and workload information [HEC05].

We have gathered and made available, as appropriate, workload data or traces from the following science codes, so far:

- S3D
- CTH
- Alegra
- PMEMD
- MILC
- MADbench
- GTC
- Paratec
- CHOMBO
- FLASH-IO
- RAGE
- NWChem
- WRF
- VASP
- MOLPRO
- ESP

We have also gathered workload, file systems statistics, and failure data from the following, so far:

- National Energy Research Scientific Computing Center
- Pacific Northwest National Laboratory
- Los Alamos National Laboratory
- Sandia National Laboratory
- Pittsburgh Supercomputer Center
- Cray Inc.
- Ask.com
- One anonymous supercomputer center and two anonymous corporate data centers

3.2 Performance Characterization

Studying the failure behavior of computing systems is critical for the effective design and implementation of petascale data storage resources. Equally important is to understand how storage systems behave when operating "normally." As part of its data collection and release subproject, several PDSI researchers collected data about I/O behavior, with a focus on applications of interest to the DOE Office of Science running on large-scale, high-performance computing systems. These days, such systems usually consist of a large number of nodes. Applications are run on a subset of the nodes called compute nodes. Applications access storage via a parallel file system such as Lustre, PanFS, PVFS, or GPFS deployed on another set of system nodes that provide system services. Application processes running on compute nodes, via systems software operating on their behalf, issue requests to the parallel file system servers running on service nodes.

There are many perspectives from which one may study this problem. One perspective focuses on the behavior of the entities generating the workload for the storage system; these are usually one or more applications running on the system but may also be middleware or systems software. Another perspective focuses on the behavior and characteristics of the storage system hardware and software itself. Another facet of this problem involves the different time scales across storage systems operate. It is valuable to study not only the system's dynamic behavior as an application runs, but because of the persistent nature of storage, it is also important to understand its behavior over longer periods of time. Each of these perspectives provides valuable information about the requirements and capabilities of a storage system. In this section, we describe the ways that PDSI researchers facilitated investigation into these perspectives through data collection, analysis, and release.

3.2.1 Application-Side Perspective

To develop and optimize an effective petascale data storage system, researchers need to know not only the capabilities and failure characteristics of storage system hardware and software, but also the demands that current and future applications will place on that storage system. Collecting and releasing data on application-side I/O and storage demands enables researchers and vendors to ensure the storage systems they propose, design, and implement will meet application needs.

SNL has collected and released event traces for scientific applications of the S3D turbulent combustion simulation running at large scale. The SNL event tracing method captures program I/O behavior at the Virtual File System (VFS) level within the system software running on the target system's compute nodes. LANL researchers have also developed event tracing tools and collected and released event traces that capture kernel-level behavior on a system's compute nodes as well as application-level behavior. Researchers at PNNL are using these trace tools to collect event traces of applications commonly run on PNNL high performance computing systems, including the NWChem computational chemistry simulation code and the WRF weather research and forecasting model. PNNL also investigated novel 3D visualization techniques for displaying event trace data (see Figure 1). The volume of trace data collected from large-scale application runs can be massive, and such visualizations can be effective in exposing interesting system behavior from this large data volume.

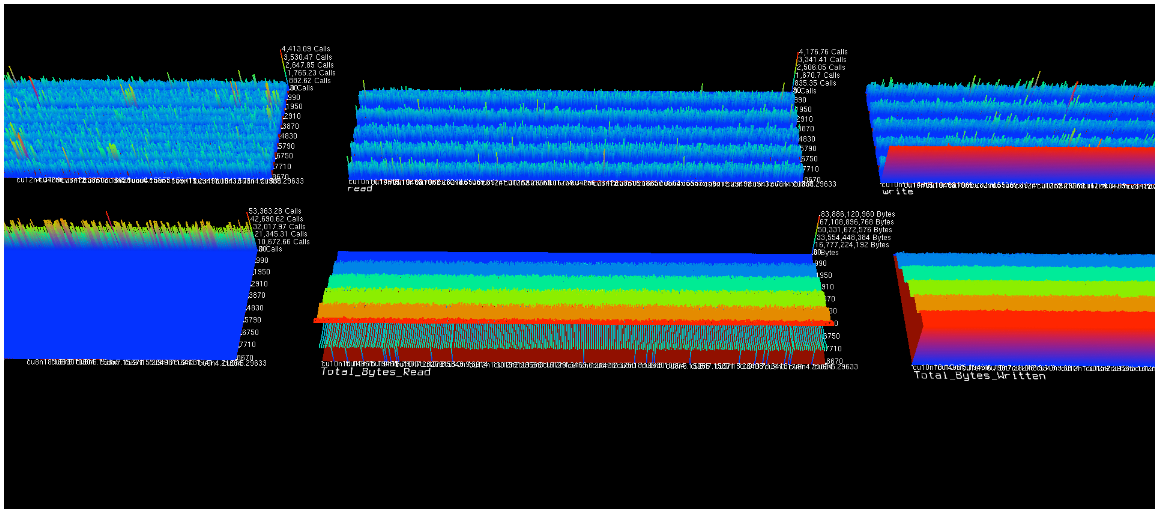
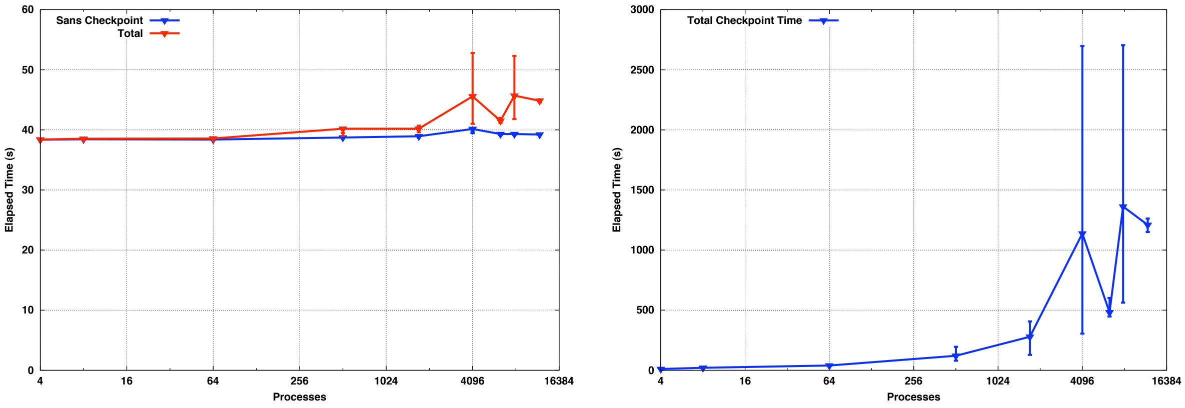


Figure 1: Example 3D visualization of PNNL event trace data. Within the figure, individual displays show number of calls to I/O functions and I/O data volume.

As part of a collaboration between PDSI and the SciDAC Performance Engineering Research Institute (PERI), PDSI researchers at both NERSC and ORNL collected event traces and producing characterizations of the I/O behavior of the applications chosen by PERI for its Tiger Team activity, including the Parallel Ocean Program (POP), the S3D turbulent combustion model (e.g., see Figure 2), the FLASH astrophysics simulation, and the GTC fusion simulation. ORNL's data collection approach complements the SNL approach by instrumenting the application executable at the end-user and middleware levels, allowing researchers to identify opportunities for optimization between the I/O operations as expressed in the end-user code and the I/O operations as seen by the kernel running on the target system's compute nodes. Within the PDSI/PERI collaboration, one use for this data is as a workload description in performance predictions.



(a)

(b)

Figure 2: Time spent performing checkpoint I/O for S3D, c2h4 problem with weak scaling. Left plot (a) shows measured time for 10 timesteps and 1 checkpoint, right plot (b) shows predicted time spent checkpointing in a 12-hour run.

3.2.2 Systems-Side Perspective

Application-side event traces and I/O characterizations focus on application demand, often limited to a single application. However, a storage system must service all applications simultaneously running on a system. Because those applications are usually not designed to make coordinated accesses to the storage system, contention at file system servers is an important aspect that can limit the storage system's effectiveness. In addition to studying application-side demand, PDSI researchers also studied storage system behavior from the server-side perspective.

Several PDSI researchers collecting application-side I/O event traces also worked toward collecting server-side data about dynamic I/O behavior. However, as with application-side event trace data, the performance data volume produced when tracing the file system servers for a large-scale system may be prohibitively large if it is monitored for a long time. However, due to its purpose, a storage system is by nature long-lived compared to the run of an individual application, and it is valuable to collect information about the storage system over longer time scales. One way that PDSI researchers investigated storage behavior over this longer time scale was by observing the static state of file systems after a long period of use. CMU and LANL researchers have developed tools for collecting a static survey of file system statistics, such as the number of files and directories in a file system and the distribution of sizes for those files (e.g., see Figure 3). PDSI is hosting a publicly accessible repository of survey results. Currently, there are nineteen survey results available in this repository, most submitted by PDSI researchers for use by the storage community.

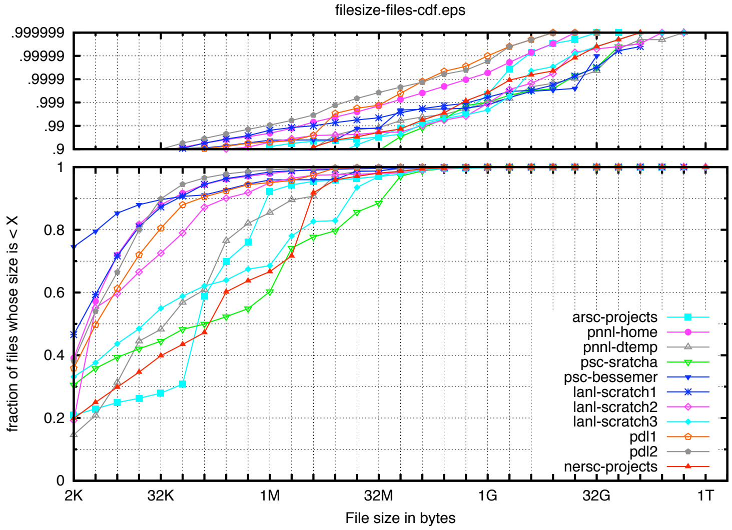


Figure 3: CDF of file sizes across eleven non-archival file systems. [Dayal-08]

3.2.3 Data Release Support

In addition to data collected and released by PDSI researchers themselves, PDSI also facilitated the distribution of data for others. For example, UCSC hosts a repository for I/O-related event traces and tools. In particular, the UCSC repository provides a mirror of tracing tools and event traces for the Storage Networking Industry Assocation (SNIA) Input/Output Traces, Tools, and Analysis Technical Working Group, facilitating wider distribution of those traces to the research community.

3.3 Failure Characterization

3.3.1 Reliability, usage, and error log data collection

As the number of components in large-scale computing systems increases, studying the ways that those components fail is increasingly important. The collection, analysis, and release of data regarding failures in storage systems was an early, high profile success for PDSI researchers. Activity in this direction was sparked when PDSI researchers from LANL secured the release of data regarding failures in twenty-two LANL cluster computer systems over a period of nine years. [LANL-FAILURE-07] The scope of this data was unprecedented, providing a window into the behavior of high performance computing systems previously unavailable to researchers in academia and industry.

The LANL failure data, along with failure data gathered from several Internet service sites, was first analyzed within PDSI by researchers from CMU. [Schroeder-07] The results of this analysis challenged several well-established rules of thumb held by hardware vendors and the storage research community. For instance, a "bathtub model" is commonly assumed for the pattern of disk drives over time. Under this model, a large number of failures will be observed when a collection of drives is first deployed. This period of "infant mortality" is followed by a stable period with a low failure rate for the drives' nominal lifetimes. As the drives near end-of-life, the number of failures rises suddenly and dramatically. In contrast to this widely held failure model, the CMU analysis indicated that for the systems studied the observed failure patterns did not exhibit a significant infant mortality period nor a period of stable replacement rates during the drives normal lifetime. Instead, drive replacement rates grew steadily with deployment age. Also, the failure data showed similar replacement rates for enterprise- and desktop-class drives, challenging another widely-held belief regarding the superiority of enterprise-class drives from a failure perspective. A paper describing this analysis received a prestigious Best Paper award at the 2007 File and Storage Technologies conference (FAST07), and its conclusions continue to have significant impact throughout the storage vendor and research communities.

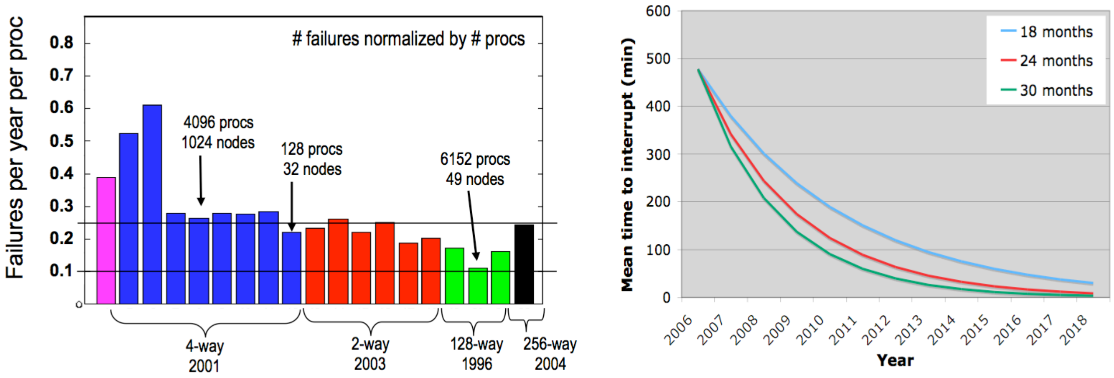
3.3.2 Shared repository for dependability data

Building on the success of the CMU analysis, PDSI has continued to shepherd the collection and release of failure data for the storage research community. CMU researchers initiated, and continue to manage, the Computer Failure Data Repository [CFDR] as a publicly-accessible repository of failure data sets. The repository currently hosts twelve data sets recording failure data from systems at high performance computing centers and commercial data services companies. The repository includes several data sets collected and submitted by PDSI researchers: the original LANL failure data set, data sets from LANL and PNNL recording hardware failures of high performance computing systems at those sites, and a data set recording I/O-specific data from several NERSC systems.

3.3.3 Dependability analysis, modeling and prediction

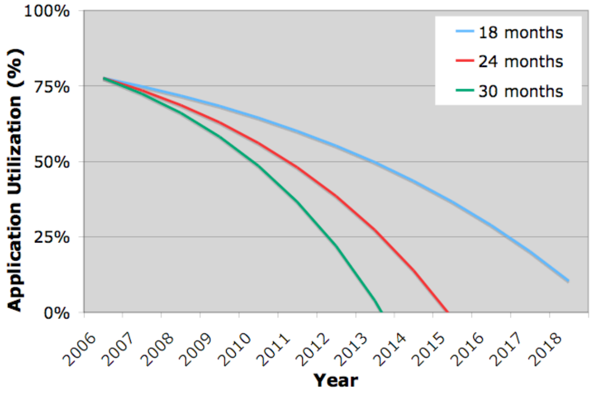
Using the first release of failure data, the decade of application interruptions in the supercomputers at Los Alamos National Laboratory, we characterized interruptions as a function of system size [Schroeder-SciDAC07]. Although this does not yield a perfect model, the best simple model suggests that the number of interrupts is linear in the number of processor chips in the cluster, regardless of the organization of processors per operating system, as shown in the following figure. With this model we project the trend for mean time to interrupt in future systems, assuming these systems continue to track the trends of top500.org: the largest computers increase aggregate speed by 100% per year. In this model we project that the per chip performance grows at best at Moore's Law, doubling every 18 months, and perhaps doubling more slowly, say doubling every 24 or 30 months, because the switch to multi-cores rather than faster cores may not increase aggregate speed as fast. Moreover we use an optimistic 0.1 interrupts per year per chip and baseline the size of the system to 1 PFLOP in 2008.

Figure 4



This model for mean time to interrupt is greatly concerning, as the time between interrupts may drop to as little as a few minutes as we approach the exascale era. For the largest applications this would demand much more frequent checkpoints in order to make forward progress. If these machines are balanced, meaning that storage bandwidth has grown linearly with compute speed, as has total memory size, then the increasing frequency of taking checkpoints will consume more and more of the machines resources, leaving less for applications, and the effective application utilization may cross under 50% before 2014, well before exascale is achieved. To emphasize this problem further, consider the cost of the storage system in these computers. Because disk bandwidth grows at only about 20% per year, scaling storage bandwidth at 100% per year, the balanced system growth rate, means the number of disks is growing at about 67% per year, much faster than the number of chips in the system, so a larger and larger fraction of the system's total cost goes into storage. This is unlikely to happen. Worse, to solve the dropping effective application utilization problem with faster storage bandwidth would require the number of disks to grow at over 130 % per year, at a rate of growth of cost that is highly unlikely, even if the disk failure rates resulting could be coped with by advanced RAID systems.

Figure 5



Of course this problem does not happen to any application that does not need large fractions of physical memory or total CPU chips, as the rate of failures for these smaller applications may not grow as fast and checkpoints might be storable in the memory of the nodes instead of on disk.

For the very largest applications, the choices are fewer. Specialized checkpoint devices perhaps, provided that the cost of these devices is less than a storage system growing fast enough to absorb the more frequent checkpoints. Certainly if the application authors can "compress" the storage footprint of a chekpoint better each year, that is by about 25-50% more effective compression each year, then the problem goes away. Finally, if we accept that the effective utilization of the machine is going to drop to 50% another option is available – process pairs – run two copies of every computation so that the failure of any one node does not make its state unavailable because there is another copy. This solution can dramatically reduce the need for taking checkpoints to the rate they are needed for visualization and steering, or perhaps as low as the rate of interrupts. In this latter case we take a checkpoint in the surviving copy of a computation after a failure has occurred elsewhere.

Checkpoint-restart fault-tolerance is clearly under a lot of pressure in extreme scale computing. Our PLFS project addresses a harder problem not covered in this model, checkpointing AMR codes into a single file. The PDSI team and many others worked on better fault-tolerance tools and mechanisms for the duration of the PDSI project.

4 Exploration

4.1 Summary

Petascale computing places very high demands on storage systems; current approaches to storage scalability are insufficient to meet these demands. We must focus on true scalability rather than simply targeting "peta" scalability to avoid facing this challenge again as scientific computing demands continue to grow.

PDSI has made great strides in addressing issues with petascale file systems and storage for high-end computing. Our researchers have produced near-term successes as well as foundational research for longer-term solutions to address long-standing needs for file and storage in support of scientific computation..

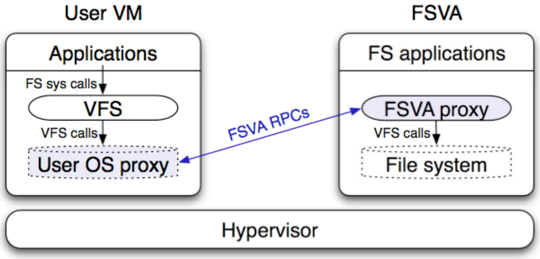
4.2 Explorations

4.2.1 Deployment/Maintenance

Reducing File System Porting Churn

Parallel file systems achieve higher performance than NFS, for example, by, in part, putting unique file system client code into client operating systems. While this enhances performance it also means that any change in the file system client code or client operating system code requires porting effort, and induces delays in accepting these changes into the user's site until file system vendors complete this porting. GPFS, PanFS, PVFS, and even OpenAFS have reported large amounts of their effort, and delay for their customers, are caused by porting to changes in the host operating system, mostly without finding any difficult changes or bugs. With the increasing amount of hardware support for virtual machines we envision that before long even HEC systems may be willing to run virtual machine monitors for the transparent job migration and management it enables. Our approach to reducing the porting costs and delays is to locate the real file system client code in a virtual machine with a stable unchanging operating system, then allow the application and site administrator to choose and advance the application operating system independently [Abd-El-Malek-PDL08-106, PDL-09-102]. Of course there is a forwarding client still needed in the application operating system, but this is simpler and common to many different client file systems, so we believe that Linux is much more likely to adopt this forwarding code. There is a performance degradation for the process switching, however, with shared memory tricks common in virtual machines, we hope that this need not slow down applications significantly.

Figure 6



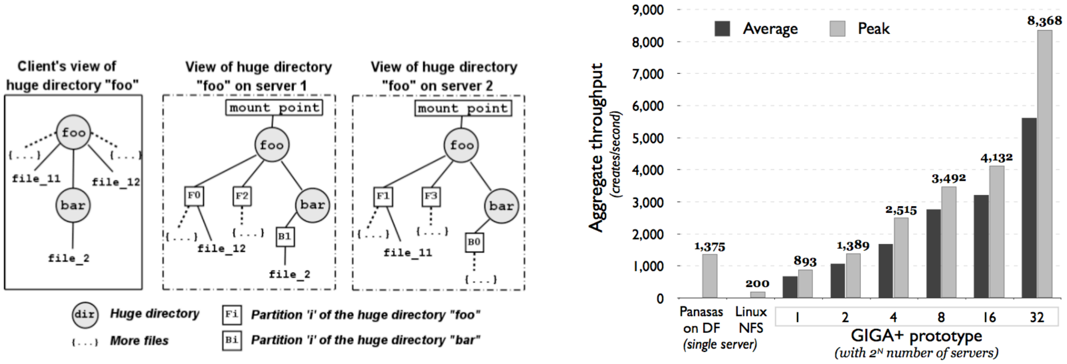
4.2.2 Metadata

We have improved support for metadata in petascale file systems on several fronts. Our research into designs for directories that can support millions of files is having a near-term impact on improving file system performance, and our explorations of new approaches to gathering, indexing, and searching file system metadata will help provide better solutions for managing petabyte-scale data in the middle to longer term.

<u>Giga+ (Hashing for Massive Directories)</u>

While bandwidth in HEC parallel file systems has in many cases been able to scale with total compute speed, metadata operations have not similarly scaled. In particular, concurrent creation of many small files in the same directory by different clients does not scale in production parallel file systems because either one server does all the work, or cache consistency protocols serialize changes in the directory. Research ideas in this area hash files into buckets on different servers, but suffer serializations in the process of growing small directories to large directories by rehashing and moving entries. CMU has developed GigaPlus, a scalable hash-partitioned parallel directory system that achieves fully parallel rehashing and allows client caches to go stale and be corrected only as needed with a minimal number of additional server messages [Patil08-PDL08-110]. We have experimented an with implementation in PVFS2, with assistance from researchers in the SciDAC SDM center, and with an implementation using the FUSE user level file system framework (see the bucket storage structure in the underlying parallel file system in the left figure) [Hase08-PDL08-107]. Performance is significantly faster than some production parallel file systems. OrangeFS, a vendor for PVFS, has integrated a version of Giga+ for HPC production use.

Figure 7



Local representation of huge directory in Giga+ prototype **Scale and performance of Giga+ using UCAR Metarates benchmark.**

<u>Content Indexing</u>

Part of our effort in improving file system metadata is the design and evaluation of different approaches to building scalable indexes that handle both file metadata searches and content-based searches. Using a partitioned approach to metadata indexing, we have developed a system that is 10–1000 times faster than existing database systems at metadata search, better allowing file system users to find locate specific files in a petabyte-scale storage system. In addition, our index requires far less space than traditional approaches built on standard databases, and is much more reliable, since failures in a portion of the index only require that portion to be rebuilt, avoiding a scan of the entire file system to rebuild a corrupt index.

Faceted Search

Another research focus is the development of metadata-based faceted search techniques to help users navigate and find valuable information in petascale file systems. In particular, we investigated techniques to automatically tailor the faceted search interface to individual users, so that users can easily view and search the relatively small part of the file system that is the most relevant for them. We have proposed a probabilistic information retrieval framework to achieve personalized collaborative faceted search to optimize user utility of the search interface. An inexpensive method to measure the potential effectiveness of different search interface is also proposed. This evaluation method involves using real world user data to generate simulations of user interactions on the search interface being tested and measuring the interface's expected utility to users.

A collaboration between LLNL and UCSC explored a path-based query language to name files in a metadata-rich file system design (MRFS) built on the foundation of POSIX in which files, their associated metadata, and relationships among files are all first class, intrinsically inter-related objects. We designed a path-based query language called QUASAR to integrate queries into file paths in MRFS, allowing the file system to perform optimizations to reduce the time consumed by operations such as reading query-derived directories, and examine characteristics of files in these directories. To evaluate the MRFS design we conducted realistic case studies in data mining environments at LLNL. Initial results suggested that MRFS dramatically reduces complexity.

Use of Solid State Storage for Metadata

A fundamental issue with scalable metadata handling is that its performance is limited by disk performance. To alleviate this performance bottleneck, we investigated the use of non-volatile RAM technologies such as NAND flash, NOR flash, and phase-change RAM. In addition to exploring designs for efficient metadata service using NVRAM, we published work on improving the reliability of NVRAM-based storage by using error-correction codes at different levels of the memory system. We believe that such an efficient, reliable NVRAM-based metadata server has the potential to dramatically improve petascale file system performance by reducing the metadata bottleneck. Panasas has since introduced HPC storage systems with NAND flash primarily used to accelerate metadata and small file access.

4.2.3 I/O Path

PLFS

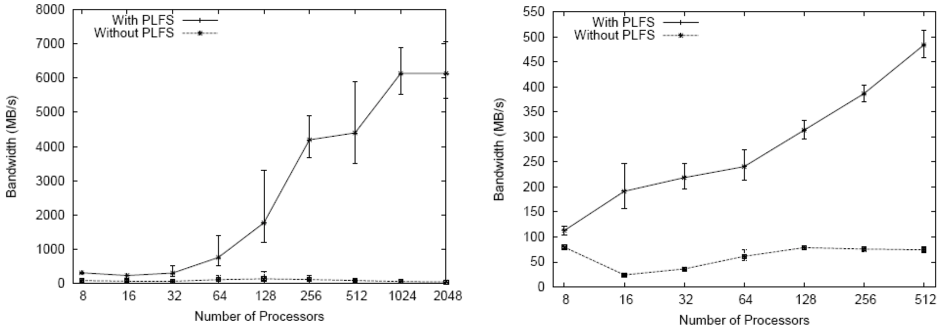
One of the biggest contributors to performance issues on SciDAC applications is the speed with which parallel applications running across thousands of machines protect themselves from inevitable failures by checkpointing: saving their state to secure, non-volatile storage. This is done for two reasons, recovery from failure, and use in visualization and other post-run analyses. For many applications, it is logically most convenient and efficient for them to save this state into a single file into which they all concurrently write; additionally, the writes are often small and not aligned with file system boundaries. Unfortunately for these applications, their preferred data layout results in pathologically poor performance from the underlying file system since it is optimized for large, aligned writes. To address this fundamental mismatch, we have developed a Parallel Log-structured File System, PLFS, which is positioned between the applications and the underlying file system and remaps the application's preferred data layout into one that is optimized for the underlying file system. Through testing at LANL, we have seen that this layer of indirection and reorganization can reduce application checkpoint time by an order of magnitude. We expect that PLFS can improve the checkpoint bandwidth for any large parallel application that does checkpoint

IO to a single file. The expected improvement is especially large for those applications doing unaligned or non-linear IO such as those applications using data formatting libraries such as NetCDF and HDF5. Because this library is interposed between the application and the underlying file system, it is straightforward to port it for use in different environments, requires NO application changes, and helps PanFS, Lustre, and GPFS file system performance for shared file check-pointing applications

The following figures demonstrate the enormous checkpoint performance gains on SciDAC applications:

Figure 8

FLASH Checkpoint I/O Chombo Checkpoint I/O



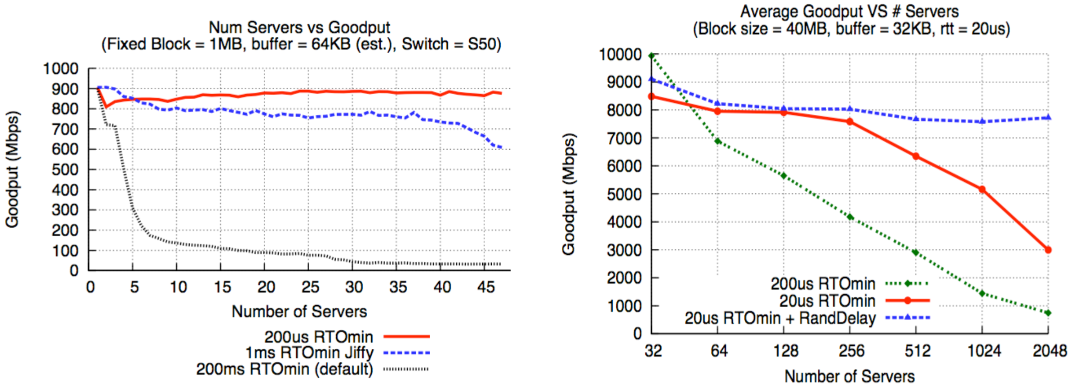
Parallel Layout

Another factor in SciDAC I/O performance is the different placement strategies used by parallel filesystems to select storage nodes for chunks of data. We used trace-driven simulation to compare the placement strategies of Ceph, PanFS, and PVFS under different workloads. We chose these file systems because their placement strategies are sufficiently different, and information about their placement strategies is readily available to us. An important contribution of this work is the construction of the simulator itself: for the simulator to provide performance results consistent with the real system, it requires a model of data placement that abstracts over the details of different file systems. Such a simulator can be used to improve workload-specific data placement, workload balancing, and quality-of-service guarantees for every parallel file system. We validated the data placement simulator using a variety of workloads from high-end scientific computing and commercial enterprise environments, and worked closely with the PLFS team at LANL to improve PLFS data placement.

Storage Area Networking

Some HEC storage servers can deliver data on Ethernet networks using TCP/IP protocols. Unfortunately synchronized sending from many IO servers to any one client causes “INCAST,” in which switch output buffers are overwhelmed and the resulting repeated timeouts of 200 msec causes the network throughput to be crushed [PhanishayeeFAST08]. We have fixed this problem by lowering minimum retransmission timeouts, previously thought to be unsafe and inefficient, because the wide area traffic that might see unsafe congestion will never try to lower its timeout below the old minimum, and the implementation uses direct timer control rather than clock interrupt logic (Linux’ high resolution timers feature) [Vasudevan09-PDL09-101]. The following two figures show synchronized reading in 1GE networks from up to 47 senders to one client, and in 10GE networks from up to 2048 senders to one client, which needs some randomization in the timeout as well as a low minimum. We have demonstrated the effectiveness of a simple change to 1 msec minimum timeout on PanFS systems and are persuading Linux to make the change to high resolution timers.

Figure 9

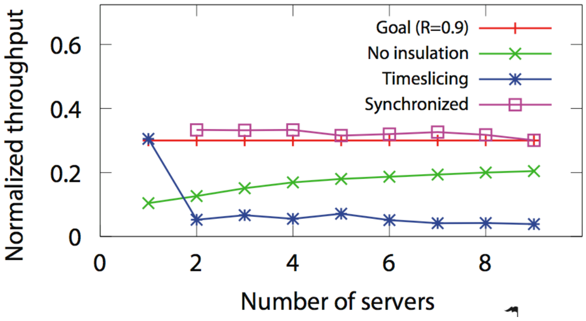


4.2.4 Management and security

Performance Isolation in Shared Storage

High performance storage is typically designed for a single, dominant parallel application. When multiple different parallel jobs are heavily using the same storage cluster at the same time, we would like to see the best case performance divided into two equal amounts delivered to each job, but often we see less fair sharing and much less total work getting done because of inefficient use of the disks, for example. Our key approach in countering this problem is to timeslice the disks, and assign jobs slices of the disk time, allowing multiple accesses from one job without long seeks to the data being used by the other job [Wachs-FAST07]. A system called Argon is providing this in the Ursa Minor object store. A job doing many small disk accesses cannot degrade the performance of another job doing large sequential accesses beyond taking a "share" of the disk time plus a small "guard band," typically less than 10% of the expected share of total disk performance. In the parallel server case lack of coordinated scheduling of timeslices can cause a further slowdown because the client waits for the last server before issuing its next request, hurting performance worse than in the uninsulated case as shown in the experiment below. Our approach is to co-schedule slices on each server, delivering about 90% of the best case performance [Wachs08-PDL08-113].

Figure 10



Scalable Security and Quota

UCSC completed a prototype implementation of a scalable security system for Ceph, a petascale storage system developed at UCSC, that facilitates strong authentication and authorization for a system with files spread across thousands of object-based storage devices. Our approach imposes very little overhead for HEC workloads; experiments on our small-scale Ceph prototype show performance degradation of at most 6–7% on workloads with shared files and shared disks, with typical overheads averaging 1–2%. Our approach integrates well with the proposed POSIX changes to support group opens, allowing them to be secured as strongly as more traditional "single node" open requests. Our approach was adapted to Hadoop, an open-source distributed computing framework.

UCSC also implemented a quota system for Ceph, allowing the system to securely track storage usage and enabling per-user storage utilization limits. This functionality is necessary for petascale I/O systems because it facilitates efficient sharing of a large storage pool between multiple users and allows the file system to better manage free storage. The policies for the quota system, which are cryptographically protected, are entirely user-controlled; thus, a system administrator can set quotas as high or as low as she wants while allowing the system itself to securely enforce the policies.

Power Management

NERSC and UCSC have explored several related issues in long-term petascale storage: usage patterns, reliability, efficiency, and durability. We investigated the use of low-power disk-based archives as a replacement for tape-based systems, providing several advantages for exabyte-scale archives: low-power usage, simpler storage evolution, high reliability, and the ability to quickly search a very large archive. These issues directly impact the ability of petascale storage users to maintain and search collections of data generated by high-performance applications. Researchers at several Department of Energy laboratories have expressed interest in using this approach for their archival storage.

Lengthening MTTI and Improving Reliability for Storage Systems

NERSC completed an analysis of component failure and system availability of NERSC computing resources. The results helped to identify problems with the HPSS production system configuration that reduced tape mounts and efficient use of tape drives that should lengthen tape drive life at NERSC. HPSS results showed that the predominant reason for HPSS outages was weekly scheduled downtime for proactive maintenance. The group is attempting to lengthen our MTTI for the storage systems by condensing multiple outages into one. Based on experience with the tape monitoring appliance, we designed and developed a prototype for a tape library monitoring application. Feedback from a demonstration to the NERSC Mass Storage Group for use with their HPSS archival storage system was very positive, and there are plans to expand the tool for use in their production environment. The application is web-based and allows storage staff to monitor tape drive and tape volume activities by providing charts of daily mounts for each drive, load frequency of tape volumes, drive utilization, time spent in drive, and other metrics. The application is readily portable to other platforms, allowing its deployment at other SciDAC sites.

The PDSI has also been developing expertise with improving storage system reliability, since recent studies using data from LANL and other have shown that drive failure is an important contributor to file system unavailability. As a result, performance/availability tradeoffs from replication have become a critical factor in long-running, compute intensive, write-mostly applications. To better understand these tradeoffs, researchers at Michigan and UCSC have developed models and tools to predict application server utilization and reliability for a given storage replication strategy. Using a discrete event simulation model, probability distributions for storage system failure and correlated failure, and characterizations of application I/O intensity and checkpoint interval, we are able to identify appropriate replication strategies to optimize application server utilization and storage system reliability.

4.2.5 Scaling

Paravirtualization

We have explored the use of paravirtualization to improve reliability at scale of petascale applications. This approach runs application processes in a virtual environment so that processes on a failed node can simply be restarted on a different node. This approach removes the need for application-specific failure code, and allows the parallel computation to survive the node failure with no ill effects. To test this approach, we built a PDSI testing cluster at PNNL using IA64 running on Xen with Lustre; this system was deployed on the Mpp2 Supercomputer housed in the EMSL User facility at PNNL. We performed wide-striped IO tests and tests using PVFS to build a large scale test system.

4.2.6 Performance

I/O Performance Modeling

ORNL characterized the I/O demands of several application programs, including the S3D turbulent combustion simulation program and the Parallel Ocean Program (POP). We selected S3D because it was one of the three applications chosen to be part of the SciDAC2 Performance Engineering Research Institute (PERI) performance measurement and modeling activity. We presented a summary of our characterization to the PERI project team in September 2008. With PERI, we developed a performance model of S3D I/O, and characterized other applications chosen for the PERI performance modeling effort.

Automatic Diagnosis of Performance Problems in a Parallel File System

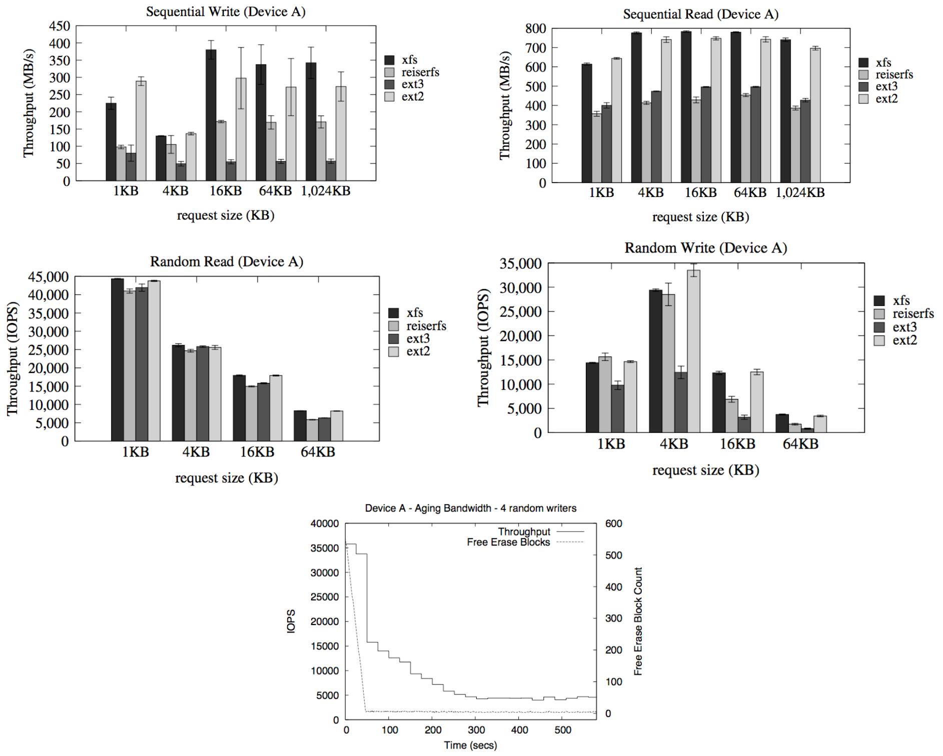
Administrators are frequently face with a user's concern about slow performance in a parallel file system. A lot of time can be spent eliminating possible sources of the problem before a problem is identified and action considered. Towards our larger agenda of automated management of parallel storage, we have experimented with automated diagnosis of typical field problems in the PVFS2 parallel file system. Our approach is to assert that problems are likely manifest themselves as rare behavior, especially different behavior than other servers in the parallel system. We apply our techniques to unmodified systems using only commonly available operating system monitoring, such as disk, CPU and network throughput and latency, and build models of common behavior using which we look for significant deviations, an approach working for us in internet service cluster experiments [Bare08-PDL08-104]. Our testing with the iozone benchmark running on a small (20 server) PVFS2 cluster and injected faults (rogue "hog" processes, blocked/lossy resources) showed at least 66% correct identification of a server suffering under an injected fault and essentially no falsely indicated servers.

A second approach to diagnosing performance problems is based on gathering tagged interprocess communication chains, identifying the end-to-end flow of the work of a client request, and comparing end-to-end latency along paths going to different servers. Applied to an experimental object store, Ursa Minor, this tool is primarily used to help developers identify code that is causing high variance in per-server performance, to enable code improvements. Tested with fault injection, metadata prefetching problems were identified and improved in an NFS server built on the object store.

Flash-enhanced Storage

Solid-state disk, based today on Flash technology, is becoming accepted as reliable enough for use in high-performance systems. This technology is highly variable today because an embedded controller is needed for leveling the wear on each flash page to avoid early wear out of some pages, and the behavior of this embedded controller is complex and rapidly changing. However, the importance to future systems is too large to make for the technology to stabilize and we have begun experimentation. Notice in the following graphs: 1) bandwidths are higher than disks, and much higher for reading (although bandwidth per dollar is not better than disk, 2) random read throughput is phenomenally higher than magnetic disks (which are closer to 100 IOPS), 3) random writes are significantly lower than random reads, and worse for sizes smaller than 4 KB, 4) different file system codes can have large differences on performance in flash probably because the magnetic disk model used by these codes is quite wrong, 5) sustained random writing performance is only good for a short time because the pre-erased page pool becomes depleted and the true cost of random writes shows through as 10 times slower [Wish09].

Figure 11



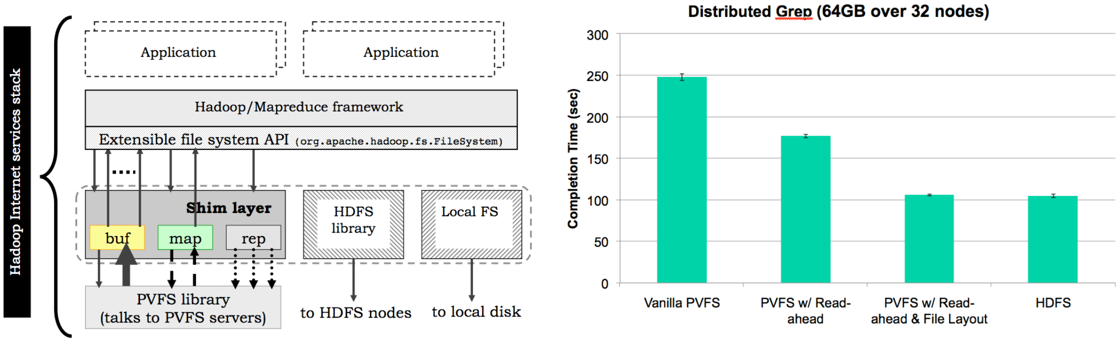
4.2.7 *Leveraging Data Intensive/Cloud Computing*

Applying Parallel File Systems to Cloud Computing

One of the core problems for the future generations of HEC storage systems is the need to invest significant research and development effort into growing bandwidth at 100% per year using disks whose bandwidth is growing at only 20% per year, and growing metadata throughput at all given that disks are basically not increasing accesses per second per disk. One way to lower the cost that the science community carries for all of the storage development effort is to spread the cost over more markets; that is, use HEC parallel file systems in non-HEC systems with significant resources. To this end we have been looking at using parallel file systems in internet services or cloud systems. The cloud systems are comparable in size to HEC systems, have healthy revenue streams and a lot of attention in the open source and academic communities.

To better understand the applicability of HEC parallel file systems in cloud uses we replaced the HDFS file system in the Hadoop software suite with PVFS2, a parallel file system maintained by the SciDAC SDM Center. Hadoop has a pluggable interface for HDFS, so we inserted a small shim that called into the PVFS client library, as shown in the figure below. No change was made in Hadoop or PVFS2, but the shim does readahead like the standard IO libraries and replication to match HDFS' three copies of every file on non-RAIDed disks. Unfortunately, the simplest shim caused Hadoop-on-PVFS to execute a large text search more than twice as slowly as the native Hadoop-on-HDFS. Fortunately with a little tuning of the readahead policy in the shim, a large improvement resulted. However, to get to full speed we need to expose to Hadoop the client addresses for the PVFS servers that contain the three copies of every file, so that Hadoop can include in its load balancing the preference for executing portions of the work on the machine that hosts the data. Fortunately, again, all parallel file systems know placement and PVFS had already exposed this information in its extended attributes. The result is that PVFS, with our shim, could be used as an alternative to HDFS in the Hadoop suite, provided that the replication was supported with failure recovery and reconstruction.

Figure 12



5 Late Project Accomplishments

Dissemination

PDSI adjusted its efforts to a higher prioritization on outreach and education. Supporting the HECFSIO national coordination of R&D including the all important second HECURA FSIO NSF solicitation was central for late stage PDSI outreach, as was following up and integrating that work into the national R&D portfolio. As the most popular workshop series in recent Supercomputing history, providing the PDSW annual workshops at SuperComputing was another vital task.

Data Collection

Data collection was an enormously important part of the PDSI mission and an area in which we changed the research community fundamentally. Continued research into failure modeling is important work as more data becomes available to lead the way for the community to round out our knowledge in this area. A few new data sets were made available at the PDSI data release site.

As always, all data, tools, models, and analysis were made publicly available as much as possible to engage as much of the research community as possible.

Exploration

LANL and CMU made progress on hardening the PLFS checkpoint acceleration middleware. With EMC's commitment of effort and NNSA assistance, PLFS in the post-PDSI era enjoys unrivaled levels of support. UCSC reported that PDSI research on archival storage resulted in the spinoff of a startup to commercialize the archival storage approaches developed by the project; we expect that this technology will be available commercially within two years. ORNL and LBNL, in cooperation with PERI, continued the effort put into understanding, modeling and improving the IO performance of SciDAC applications S3D, GTC, and FLASH on DOE leadership computing platforms at ORNL, ANL and LBNL. Monitoring, analysis, visualization and prediction tools were developed in cooperation with PERI's tool chain.

5.1 Carnegie Mellon University

Project 1: Petascale Data Storage Outreach

CMU participated in a review of coming magnetic disk technologies, and presented a PDSI perspective at the 2009 international magnetic recording conference. The most promising short range technology changes, shingled-writing and two-dimensional magnetic recording, TDMR, may change core characteristics of magnetic disk operation and require systems software be adapted significantly. With shingled writing a band of adjacent tracks overlap one another, to significantly increase data density with essentially today's recording heads and media. Once overlapped, however, a track cannot be updated in place, because the tracks overlapping it will be overwritten by the update. If this behavior was to be exposed to operating systems directly, there may be very low marketplace acceptance of these products. Our analysis showed how disk controller software could emulate full compliance with existing interfaces, and may be able to mask almost all performance implications as well. Specifically, while shingled-writing imposes serious change on the order that sectors must be written, it impact can be masked with software in the disk controller in much the same manner as solid state disks (SSD) mask the need to erase a block before writing any part of it. A bigger problem for this technology comes from initial designs of TDMR which called for three to five rotations minimum to read any data, so that overlapped writing crosstalk could be measured. We advocated that shingled writing be match with a read technology that requires no more than one rotation minimum to read data. This presentation was challenging for some technologists to accept, but it has been very influential on magnetic disk technology directions, and the short paper with it has been widely read.

Project 4: Protocol/API extensions for Petascale Science Requirements

Parallel file systems achieve higher performance than NFS, for example, by, in part, putting unique file system client code into client operating systems. While this enhances performance it also means that any change in the file system client code or client operating system code requires porting effort, and induces delays in accepting these changes into the user's site until file system vendors complete this porting. GPFS, PanFS, PVFS, and even OpenAFS have reported large amounts of their effort, and delay for their customers, are caused by porting to changes in the host operating system, mostly without finding any difficult changes or bugs. With the increasing amount of hardware support for virtual machines we envision that before long even HEC systems may be willing to run virtual machine monitors for the transparent job migration and management it enables. Our approach to reducing the porting costs and delays is to locate the real file system client code in a virtual machine with a stable unchanging operating system, then allow the application and site administrator to choose and advance the application operating system independently. Of course there is a forwarding client still needed in the application operating system, but this is simpler and common to many different client file systems, so we believe that Linux is much more likely to adopt this forwarding code. There is a performance degradation for the process switching, however, with shared memory tricks common in virtual machines, we hope that this need not slow down applications significantly.

Project 5: Exploration of Novel Mechanisms for Emerging Petascale Science Requirements

Solid-state disk, based today on Flash technology, is becoming accepted as reliable enough for use in high-performance systems. This technology is highly variable today because an embedded controller is needed for leveling the wear on each flash page to avoid early wear out of some pages, and the behavior of this embedded controller is complex and rapidly changing. However, the importance to future systems is too large to make for the technology to stabilize and we have begun experimentation. Experiments have shown that: 1) bandwidths are higher than disks, and much higher for reading (although bandwidth per dollar is not better than disk, 2) random read throughput is phenomenally higher than magnetic disks (which are closer to 100 IOPS), 3) random writes are significantly lower than random reads, and worse for sizes smaller than 4 KB, 4) different file system codes can have large differences on performance in flash probably because the magnetic disk model used by these codes is quite wrong, 5) sustained random writing performance is only good for a short time because the pre-erased page pool becomes depleted and the true cost of random writes shows through as 10 times slower. There is much work to be done to find the best utilization of solid state disks in HEC systems.

One of the core problems for the future generations of HEC storage systems is the need to invest significant research and development effort into growing bandwidth at 100% per year using disks whose bandwidth is growing at only 20% per year, and growing metadata throughput at all given that disks are basically not increasing accesses per second per disk. One way to lower the cost that the science community carries for all of the this storage development effort is to spread the cost over more markets; that is, use HEC parallel file systems in non-HEC systems with significant resources. To this end we have been looking at using parallel file systems in internet services or cloud systems. The cloud systems are comparable in size to HEC systems, have healthy revenue streams and a lot of attention in the open source and academic communities.

To better understand the applicability of HEC parallel file systems in cloud uses we replaced the HDFS file system in the Hadoop software suite with PVFS2, a parallel file system maintained by the SciDAC SDM Center. Hadoop has a pluggable interface for HDFS, so we inserted a small shim that called into the PVFS client library, as shown in the figure below. No change was made in Hadoop or PVFS2, but the shim does readahead like the standard IO libraries and replication to match HDFS' three copies of every file on non-RAIDed disks. Unfortunately, the simplest shim caused Hadoop-on-PVFS to execute a large text search more than twice as slowly as the native Hadoop-on-HDFS. Fortunately with a little tuning of the readahead policy in the shim, a large improvement resulted. However, to get to full speed we need to expose to Hadoop the client addresses for the PVFS servers that contain the three copies of every file, so that Hadoop can include in its load balancing the preference for executing portions of the work on the machine that hosts the data. Fortunately, again, all parallel file systems know placement and PVFS had already exposed this information in its extended attributes. The result is that PVFS, with our shim, could be used as an alternative to HDFS in the Hadoop suite, provided that the replication was supported with failure recovery and reconstruction.

Project 6: Exploration of Automation for Petascale Storage System Management

High performance storage is typically designed for a single, dominant parallel application. When multiple different parallel jobs are heavily using the same storage cluster at the same time, we would like to see the best case performance divided into two equal amounts delivered to each job, but often we see less fair sharing and much less total work getting done because of inefficient use of the disks, for example. Our key approach in countering this problem is to timeslice the disks, and assign jobs slices of the disk time, allowing multiple accesses from one job without long seeks to the data being used by the other job. A CMU system called Argon is providing this in the Ursa Minor object store. A job doing many small disk accesses cannot degrade the performance of another job doing large sequential accesses beyond taking a "share" of the disk time plus a small "guard band," typically less than 10% of the expected share of total disk performance. In the parallel server case lack of coordinated scheduling of timeslices can cause a further slowdown because the client waits for the last server before issuing its next request, hurting performance worse than in the uninsulated case as shown in the experiment below. Our approach is to co-schedule slices on each server, delivering about 90% of the best case performance.

Administrators are frequently faced with a user's concern about slow performance in a parallel file system. A lot of time can be spent eliminating possible sources of the problem before a problem is identified and action considered. Towards our larger agenda of automated management of parallel storage, we have experimented with automated diagnosis of typical field problems in the PVFS2 parallel file system. Our approach is to assert that problems are likely manifest themselves as rare behavior, especially different behavior than other servers in the parallel system. We apply our techniques to unmodified systems using only commonly available operating system monitoring, such as disk, CPU and network throughput and latency, and build models of common behavior using which we look for significant deviations, an approach working for us in internet service cluster experiments. Our testing with the iozone benchmark running on a small (20 server) PVFS2 cluster and injected faults (rogue "hog" processes, blocked/lossy resources) showed at least 66% correct identification of a server suffering under an injected fault and essentially no falsely indicated servers.

A second approach to diagnosing performance problems is based on gathering tagged interprocess communication chains, identifying the end-to-end flow of the work of a client request, and comparing end-to-end latency along paths going to different servers. Applied to an experimental object store, Ursa Minor, this tool is primarily used to help developers identify code that is causing high variance in per-server performance, to enable code improvements. Tested with fault injection, metadata prefetching problems were identified and improved in an NFS server built on the object store.

5.2 Lawrence Berkeley National Labs / NERSC

In April 2009, NERSC began work on evaluating Flash storage in the HPC Center. We started by installing several different PCIe based Flash devices in a test system and performing a variety of benchmarks on the devices with different configurations and drivers. IOZone produced the most useful benchmarks and enabled us to identify problems and limitations of the two different products to vendor engineers. Upon problem resolution, we were able to determine that further research is warranted due to the varied performance of the two PCIe based Flash solutions currently available. We presented our early results at the UPC Review in July 2009 at LBNL. A more detailed presentation of our evaluation was provided at the HECFSIO workshop in August 2009.

Our latest efforts have been focused on database performance, especially with the HPSS archival storage system, and application performance mainly with several I/O bound projects out of the Joint Genome Institute (JGI). JGI identified two projects, one working to improve the performance of V-Match and the other attempting to accelerate queries using IMG, that require many core, large memory systems with enterprise level disk arrays to perform analytics on DNA sequence data.

5.2.1 NERSC Parallel HDF5 Performance Analysis Project

The HDF5 library is the third most commonly used software library package at NERSC and the DOE Scientific Discovery through Advanced Computing (SciDAC) program. It is also the most commonly used I/O library across DOE computing platforms. And, HDF5 is also a critical part of the NetCDF4 I/O library, used by the CCSM4 climate modeling code, a major source of input to the Intergovernmental Panel on Climate Change's assessment reports. Because parallel performance of HDF5 had been trailing on newer HPC platforms, especially those using the Lustre filesystem, NERSC funded and worked with the HDF Group to identify and fix performance bottlenecks that affect key codes in the DOE workload, and to incorporate those optimizations into the mainstream HDF5 code release so that the broader scientific and academic community can benefit from the work.

NERSC sponsored a workshop to assess HDF5 performance issues and identify strategies for improvement with DOE Office of Science application scientists, Cray developers, and MPI-IO developers. NERSC then initiated a collaborative effort to implement that strategy. We used a number of I/O and filesystem profiling tools, including IPM (Integrated Performance Monitoring) to get a deep understanding of the performance issues.

The resulting improvements include the following:

- Increased parallel I/O performance by up to 33 times.
- Raised performance close to the achievable peak of the underlying file system.

- Achieved 10,000 GB/s write bandwidth (for certain configurations) of both applications.
- Increased scaling up to 40,960 processors.

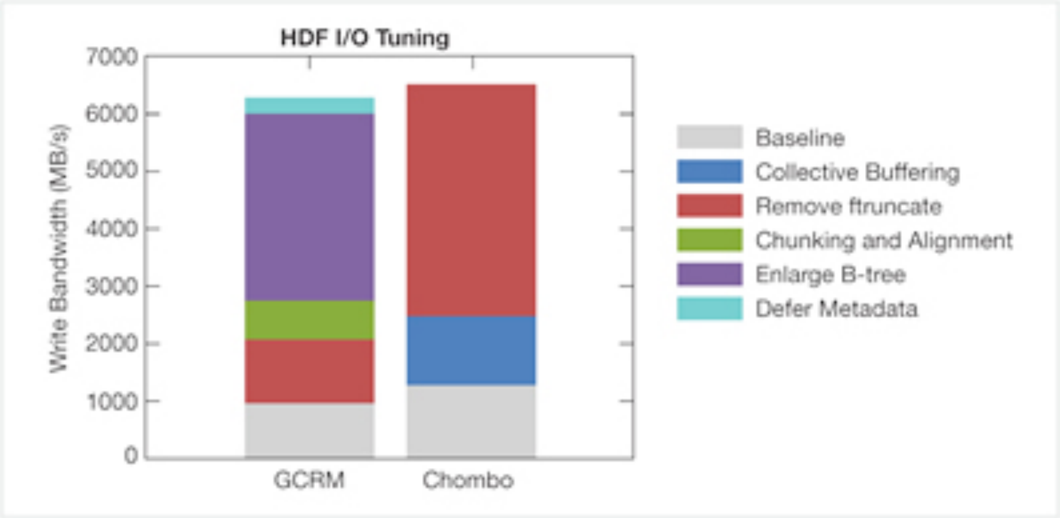


Figure 13: Cumulative benefits resulting from HDF5 performance optimizations for Chombo and the Global Cloud Resolving Model (GCRM). Peak filesystem bandwidth is 10,000 MB/s.

Figure 13 shows performance improvements HDF5 tuning work contributed to two scientific codes. The baseline (gray) is the original performance. Colored bars on top of the baseline represent performance benefits derived from incremental application of optimizations. GCRM is the Global Cloud Resolving Climate Model from Colorado State University and Chombo is the adaptive mesh refinement framework from Berkeley Lab—two very demanding, I/O intensive codes. More recent tests have achieved 10,000 GB/s write bandwidth for certain configurations of both codes.

5.2.2 NERSC FLASH I/O Evaluation

LBNL worked on enhancing our understanding of the performance characteristics of various flash devices, building upon the work described above, with a view to gaining an increased understanding of the best role for solid-state disks in HEC systems. To this end we measured the performance characteristics of five NAND flash based devices, three PCIe and two SATA attached ones. The peak bandwidth and IOPs measured using each device with the ioZone benchmark are shown in Table 1.

Table 1 Performance Characteristics of the Flash Devices

Device	Connection Type	Peak Bandwidth MB/s		I/O (4K) operations per second x10^{3}	
		Read	Write	Read	Write
Intel X25-M SATA	SATA	200	100	19.1	1.49
OCZ Colossus SATA		200	200	5.21	1.85
FusionIO ioDrive Duo	PCIe-4x	800	690	107	111
Texas Memory Systems RamSan20		700	675	143	156
Virident tachIOn	PCIe-8x	1200	1200	156	118

There are several interesting performance characteristics as compared to a regular spinning disk, and also some interesting differences between the flash devices themselves. Typically a regular SATA hard drive today can support approximately 80 MB/s or 90 IOPs for both read and write. Thus a flash device can provide a significant bandwidth and IOPS advantage over a traditional hard drive based storage system. We also observed significant differences between the bandwidth and IOPs characteristics of the SATA devices as compared to the PCI ones, albeit with increased cost. We also note that there is a fair amount of variation between the capabilities of the devices, and particular device is optimal across all four measurements.

This study was based upon more recent implementations of flash devices and shows some of the progress that has been made in alleviating the performance concerns described in the previous work. For example, the random read and write rates for the PCIe based devices evaluated here are much closer to each other than was previously observed.

The figure below shows the result of an experiment where 4K blocks were randomly written to a file that spanned 90% of the capacity of each drive for the period of 1 hour. Note that each device shows quite different behavior, which seem to depend upon how much ‘extra’ flash storage is present on each device to allow for grooming, as well as the algorithm that each uses within its translation layer and controller. The principle finding is that sustained random writing performance is now good for significant periods of time for most of the devices, which again reflects the advances in flash technology since the above analysis.

Overall therefore, as observed above, there is still much work to be done to understand the differences between various flash devices and choose the optimal device for a particular workload.

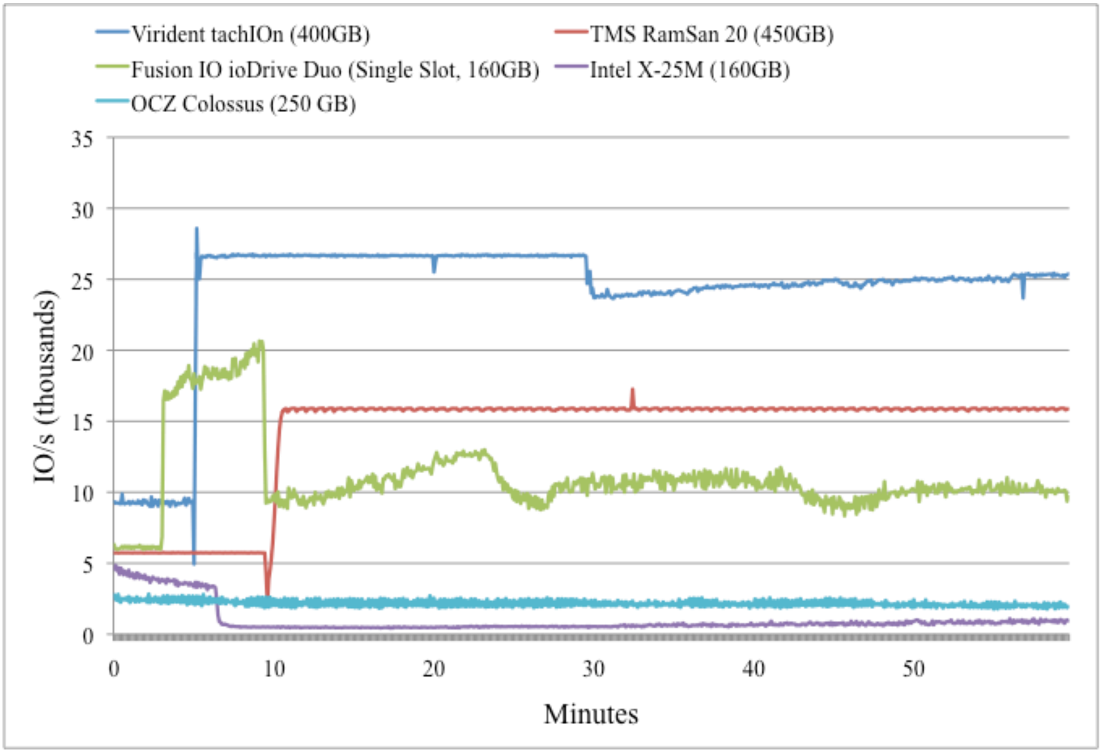


Figure 14: IOPs Performance Degradation for Flash Devices

5.2.3 NERSC Tape Media Verification Project

NERSC verified archive tape media from its original tape libraries for readability. Over 40,000 pieces of media were read from beginning to end to copy data to new media. Statistics about the success or failure of this copy were captured. The group also has a custom appliance called Archive Verify by Crossroads that performed full read verification on tape media as a separate and distinct set of statistics on the readability of data on tape. The study statistics have shown that enterprise tape media is extremely reliable (we have a 99.945% probability of being able to read 100% of the data on each tape).

Performing this exercise taught us that the appliance provides a valuable first check on bad tape media but that it often takes multiple times (3 to 5 times) reading the worst tapes in order to retrieve data. The appliance by default only reads each tape once. The appliance is useful in identifying suspect bad tapes, but will require further verification in order to narrow the extent of data availability.

Here is the summarized information from our data migration effort from 6/2009 to 3/2010. We read 100% of the data on the following enterprise tape cartridges totaling over 5 Petabytes of data stored:

– 6,859 Oracle T10KA (up to 2 yrs old)

– 9,155 Oracle 9940B (up to 8 yrs old)

– 7,806 Oracle 9840A (up to 12 yrs old)

Of those tapes, 13 tapes had data that couldn't be read. The data was in 14 files totaling less than 100GB of data. These statistics support the notion that enterprise tape remains reliable for archival storage purposes.

5.3 Los Alamos National Laboratory

LANL Data Release

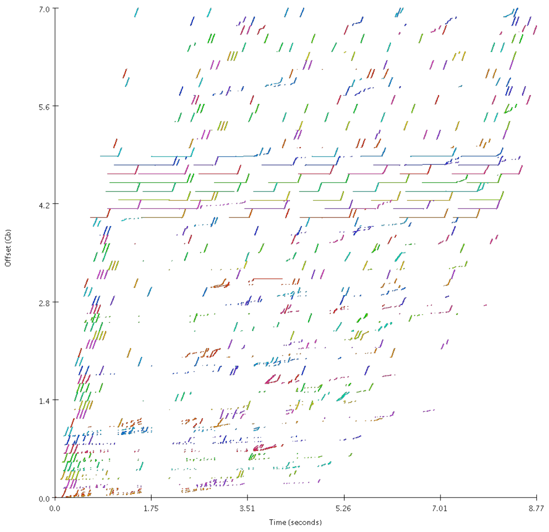
Los Alamos National Lab has an ongoing effort to release data helpful to and requested by the HPC research community. LANL has released significant amounts of trace data, file system statistics, and related software. Using the newly developed PLFS tracing mechanism (see below), LANL released almost 100 traces from seven different benchmarks and applications including three important LANL codes. In addition, LANL has released a visualization tool, Ninjat, that can be used to turn a trace of IO to a single file into images and animated movies which clearly show patterns of IO such as strided and non-strided and sequentiality. LANL continues to release file system statistics data using a tool from Carnegie Mellon University and Panasas, fsstats, for several file systems that are used by clusters that are in production and hundreds of statistics from workstation back-ups.

PLFS

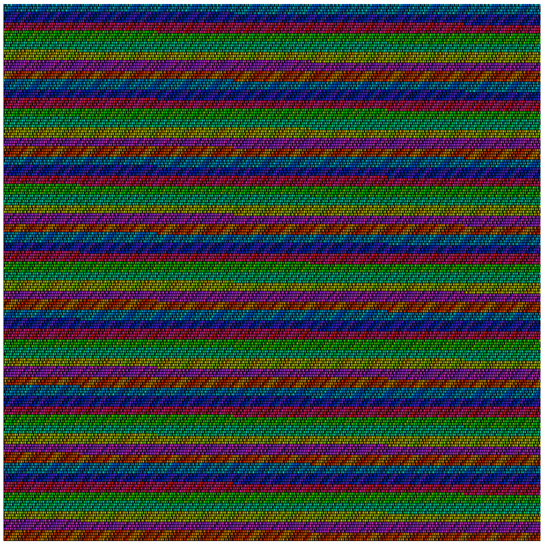
LANL developed a virtual parallel filesystem called PLFS which fundamentally addresses a decade old parallel storage challenge. PLFS transparently rearranges application IO to achieve orders of magnitude bandwidth improvements for many important applications, which use a particular pattern of IO. Most important is the transparent nature of PLFS, which allows this bandwidth improvement without requiring any additional work from the user and without forcing them to modify, in any way, their standard practices. PLFS was demonstrated to achieve these bandwidth improvements on three different parallel filesystems (PanFS, Lustre, GPFS) and with seven applications and benchmarks. PLFS was nominated as one of the three best papers at the SC09 conference.

During the past year, the HPC-5 Filesystems and I/O team has made the following contributions. We are continuing our efforts to move PLFS into production at LANL. Our final PDSI status is that we've coordinated with user groups and are working with them to start using PLFS on an initial test machine. We are working on getting PLFS installed on all LANL machines. We have also been running PLFS on Oak Ridge's Jaguar machine and are helping AWE in evaluating PLFS for their workloads. We ran their workload at LANL using PLFS and showed a large improvement with their synthetic so they are currently trying to replicate this and then show a similar speedup on their production codes.

LANL also released Ninjat, a tool for visualizing patterns of parallel concurrent writes to a shared file:



(a)



(b)

Figure 15. *These images were produced using Ninjat, a visualization tool for concurrent accesses to a single file. Ninjat was developed by a PDSI sponsored student, Calvin Loncaric, towards satisfying LANL's PDSI task of producing tools to assist in parallel IO analysis. These images were produced from traces captured by PLFS from an anonymous LANL application; both the traces and PLFS were also LANL PDSI deliverables. The image on the left shows the IO activity to a single file where each line on the graph represents a single write. The time of the right is shown on the x-axis, the offset on the y-axis, and the color of the right corresponds to the rank of the process which issued the write. The image on the right is a representation of the file as a single linear array wrapped into rows within a rectangle. This image clearly demonstrates that the pattern of IO to this particular file was an N-1 strided pattern in which each variable in the parallel simulation has a unique region in the file with the implication being that each rank in the parallel job did small, unaligned writes interleaved with writes by other ranks throughout the entire file. Although not shown here, Ninjat can dynamically draw these images in a "movie" view which is particularly useful for the image on the right, allowing the viewer to derive visual intuition into the concurrency of the parallel writing.*

5.4 Oak Ridge National Laboratory

5.4.1 Goals

Oak Ridge National Laboratory (ORNL) is a recognized leader in high-performance computing in support of open computational science. Our primary goal as part of the Office of Science's Scientific Discovery through Advanced Computing 2 (SciDAC2) Petascale Data Storage Institute (PDSI) was to understand the I/O behavior of current and future scientific applications, with special focus on their behavior on Office of Science computing resources deployed at ORNL. With this insight, we will be in position to provide guidance to application developers to make better use of the available I/O resources, to provide guidance to hardware vendors for designing systems that are more effective in servicing application I/O demands, and to provide real-world information about I/O best practices to students of both computer science and computational science.

5.4.2 Activities and Accomplishments

For this project we have developed an infrastructure for characterizing the I/O behavior of scientific applications. This infrastructure uses instrumented functions that wrap interesting library functions such as the system I/O functions read and write and relevant MPI functions. Our primary target platform remains the Cray XT system of the Leadership Computing Facility at ORNL.

Recent activities using this infrastructure involve characterization of applications of interest to the Office of Science. For instance, we had a collaboration with the SciDAC2 Performance Engineering Research Institute (PERI) for performance measurement and prediction. Whereas PERI has chosen to focus on an application's computation and communication behavior but not I/O, our PDSI team fills the gap by investigating I/O behavior. In collaboration with researchers from PERI and other projects, in 2009 we initiated an expansive performance measurement and modeling effort focusing on the Community Climate System Model (CCSM). Our intent was to model an upcoming release of the CCSM code. Although we had been given a skeleton version of the new version of the model during PDSI, the code developers had not released the complete model code in time. Therefore, our work was done with the skeleton and existing publicly available versions of the CCSM components.

In early 2009, the ORNL PDSI also initiated a project with researchers from North Carolina State University (NCSU) on scalable event tracing and replay at multiple levels of the I/O software stack. ScalaTrace is an existing event data capture and replay library that collects performance data describing MPI and MPI-IO events. To control event trace file size, ScalaTrace recognizes repetitious behavior patterns (e.g., loops) and saves information describing the pattern rather than detailed information about each event. The ORNL PDSI team augmented ScalaTrace to collect and compress data describing POSIX I/O events in addition to MPI-IO events. We also modified the ScalaTrace event trace replay mechanism to support user-defined actions instead of just regenerating MPI events. We used this capability for workload analysis and integrated it with our PERI-funded simulation-based performance prediction framework, allowing the use of ScalaTrace event trace files for specifying simulation workloads. As part of the scalable event tracing project described above, in the summer of 2009 we hosted a student from the NCSU ScalaTrace project.

In 2010, ORNL closed out PDSI I/O research in two directions: prefetching and layout-aware collective I/O. With respect to prefetching, ORNL proposed and evaluated extensions to existing prefetching techniques to support Global Multi-order Context-based (GMC) prefetching. GMC uses multi-order analysis using both local and global context to increase prefetching coverage while maintaining prefetching accuracy. ORNL also investigated the benefits of layout-aware collective I/O to parallel I/O performance. With layout-aware collective I/O, the physical layout of data in a parallel file system is exposed to the I/O middleware implementing collective I/O (in this case, a modified version of the ROMIO MPI-IO implementation and a PVFS2 parallel file system). Using this information about physical data layout, the middleware can optimize collective I/O requests. In our evaluation, the proposed layout-aware approach showed performance benefits of at least 24% for the tested benchmark workloads, with the benefit increasing as the number of processes increases.

5.4.3 Outreach Activities

The PDSI team did not develop a workshop explicitly for application developers and users. However, with members of the SciDAC2 Scientific Data Management (SDM) project, ORNL PDSI team members proposed an HPC I/O tutorial to the 2008 ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP). The tutorial was accepted, but cancelled.

5.5 Pacific Northwest National Laboratory

5.5.1 Support of Large Scale IO tracing on PNNL supercomputer

In 2009 the PNNL PDSI team deployed the LANL trace library on the Chinook supercomputer at PNNL. The system was set up so that any users of the system could make some small changes to the job submission scripts and after running the job, have a full strace or ltrace based output of the entire job. Scripts were then provided to translate these traces into a CVIEW data set that could be visualized using the CVIEW 3d interactive graphics software. The PNNL team ran applications with known performance characteristics and observed the expected behavior using the trace tools. The team also gained performance insight into numerous other applications through these application traces including NWChem, WRF/CHEM, and IOR.

5.5.2 Novel Virtualization Data Subsystems

The PNNL PDSI team investigated IOMMU virtualization technologies, and procured a new Hardware RAID subsystem that supports running User-created Virtual Machines on the Internal processors. PNNL purchased early this year a 1U server that had the required CPU and IOMMU chipsets so that we could test having raw PCI space for IO controller cards directly accessible to the Virtual Machines running on the node. Tests have shown that there is a benefit to having this technology in a VM.

5.5.3 Outreach Activities

- Parallel File System outreach: Debian, Ubuntu, Redhat 5.X, and CentOS 5.X Parallel filesystem packages for Lustre and PVFS have been maintained and available throughout the year.
- The student previously interning with PDSI at PNNL was hired as a full time researcher.

5.6 Sandia National Laboratory

High Performance System Call Tracing

Ms. Chen's S3D application I/O kernel was traced on Red Storm, at scale. This was done three times with the different I/O schemes supported. The collective and "FORTRAN" I/O traces are released to the public and available at http://www.cs.sandia.gov/Scalable_IO/SNL_Trace_Data. The previously acquired Alegra and CTH traces remain available there, of course.

The third I/O trace for S3D used the HDF5 library and was almost unusably slow. Using the traces, we repaired the S3D implementation, resulting in many orders of magnitude performance improvement, and gave the modifications back to Dr. Wei-keng Liao at Northwestern University, the original author. Since repairing the HDF5 library we have been unable to reacquire the machine for testing as other, mission related, priorities have monopolized it.

We have continued to work on the Linux version of the Red Storm tracer, albeit with little forward progress. The code collection is now in its third version and nagging locking issues remain. With our final six months funding addressed these issues and we hope to soon release the code.

The goal of this project was to develop a very low overhead system call trace facility for two operating systems; Sandia's light-weight kernel as deployed on the Cray XT series machines called Catamount, and Linux. Although by our plans funding was to be terminated approximately mid-way complete, we were able to refocus our effort and the Catamount version of the tracer was completely accomplished and the data released to the public. The refocusing occurred while the Linux version of the tracer was at the height of its development and, per the close-out plan, was pared back to bare essentials and a partially functional, not terribly useful, version of it was prepared. Another sponsor, a member of the high performance IO community, intervened and offered to fund the work through completion. The Linux tracer was completed and is in use now as a performance analysis tool by the new sponsor.

5.6.1 Outreach Activities

In addition to the non-specific outreach activities, Sandia requested funding for summer students in this contract. This year, we again hosted Matt Curry from University of Alabama at Birmingham. Matt accomplished a prototype, iSCSI targeted RAID controller. As this is application and no longer research the work was funded by another project this year. It is reported here, though, as it relates to impact due to this grant. Directly funded by this grant, Robert Cloud, another student from the University of Alabama at Birmingham, accomplished his first summer intern position at Sandia this year. Robert investigated the efficacy of compression algorithms executed within a GP-GPU environment. Robert implemented the traditional Huffman compressor, suitably modified for the GPU environment. This preliminary work displayed large-block compression performance around 250 MB/s and decompression rates at roughly twice that. Robert has crafted a paper detailing his accomplishments but has not yet submitted it for publication.

5.6.2 Close Out Activities

In our close-out plan we promised to attempt to transition the work to Northwestern University and focus on bringing the Linux development effort to a working but limited condition.

In this, we were only marginally successful. The Northwestern collaboration resulted in the paper, mentioned above, but did not prove sufficiently interesting enough to motivate Northwestern to take over the lead. We did accomplish all of the technical milestones we set for ourselves but the resulting product was sub-optimal. It was too limited, and so sensitive to the operating environment that its use in a production environment was never really viable.

Catamount is no longer in use on any production system in the world. As such, the published version of its tracer code as limited utility. As part of the project, the SYSIO library used on Catamount was significantly modified and those modifications are available in the main, default, branch of the public source tree which may be found at http://sourceforge.net/projects/libsysio/ by following the appropriate links to the tree. The tracer-specific module for that library together with the post-processing tool to interpret the trace-file format was never made available, because of its limited utility, but could be requested by contacting the former Principal Investigator of the project.

5.7 University of Michigan

5.7.1 pNFS Research and Development

pNFS is an extension to NFSv4 that helps clients overcome NFS scalability and performance barriers. Like NFS, pNFS is a client/server protocol implemented with secure and reliable remote procedure calls. pNFS departs from conventional NFS by allowing clients to access storage directly and in parallel. This helps overcome server bottlenecks inherent to NAS access methods.

Making pNFS available to petascale researchers required coordinated progress in several dimensions, a confluence of multiple processes that took nearly ten years. The protocol had to be specified and published by the IETF. Nascent implementations had to track changes in the draft specification, changes in the Linux kernel, and interoperate with one another. The process by which modifications are accepted into the Linux kernel by maintainers and developers (and ultimately by Linus Torvalds himself) itself required consensus and compromise. Finally, pNFS support needed to be provided by the major Linux distributors (Red Hat, SUSE, etc.) and storage vendors (Netapp, EMC, IBM, Microsoft, etc.) on both the engineering and product sides.

The specification requirement was met in December 2008, when the IETF NFS working group forwarded the NFSv4.1 specification, which incorporates pNFS and pNFS file layouts, to the IETF architects as a Proposed Standard as well as the (separate) specifications of pNFS object and block layouts.

Throughout the PDSI project, CITI was the key contributor to the Linux-based, open source implementation of NFSv4.1 and pNFS. Considerable effort was devoted to the process of refining the IETF specification, which underwent dozens of preliminary drafts, rebasing implementations to the latest Linux kernel, which itself changed quarterly, and to testing interoperability with other developers. At last, the Linux pNFS implementation was incorporated into the Linux mainline kernel, although this was achieved after the PDSI project ended.

Implementation

CITI is one of a handful of primary contributors to the Linux-based, open source implementation of NFSv4.1, which includes pNFS. Considerable effort was devoted to the process of rebasing implementations to the latest Linux kernel and to the final IETF draft specification.

Linux kernels are released approximately quarterly. The real focus of Linux development is in the development kernel; whenever a new kernel is released, a development follow-on is opened up within a few days. At that point, Linux maintainers (e.g., the NFS client maintainer, Trond Myklebust, or the NFS server maintainer, Bruce Fields) have the opportunity to merge in new modules and components for the next release. The remaining months are spent making the result stable and bug-free.

The elements of the Linux NFSv4.1 implementation are maintained in a private `git` tree that builds on the mainline kernel. This implementation is functional and interoperates with vendor implementation efforts (Sun, EMC, IBM, etc.), but is not yet tested at scale.

Release

Complementing the IETF NFS committee's completion of the NFSv4.1 specification, FY09 also saw the first NFSv4.1 components accepted into the mainline Linux kernel. In particular, the forward and back channels of the NFSv4.1 communication layer (called Sessions) were first accepted into the mainline kernel. FY10 saw significant revision and debugging in this code base without any major features being released by Linux. It was not until after PDSI was shutdown that Linux was satisfied with the client versions of the full featured pNFS functionality – Linux release was started in April 2011 and first completed in February 2012.

pNFS server code is in a less complete state, because vendors do not see its need or support its development. CITI developed a file metadata layout server by extending GFS2, one of two cluster file systems in the Linux kernel. Lock contention in GFS2 may limit scalability in large clusters, but CITI developed a performance test bed — an eight-node cluster that uses Linux iSCSI targets as shared storage — to continue scaling testing if subsequent need and funding is determined.

5.7.2 Outreach Activities

- Peter Honeyman organized a BoF session on HPC Storage at the FAST Conference, February 2010.
- CITI participated in the spring 2010 pNFS/NFSv4.1 interoperability workshop (Connect-a-thon), held in Santa Clara, CA.
- CITI organized and hosted the summer 2010 pNFS/NFSv4.1 interoperability workshop (Bake-a-thon), attended by dozens of technologists from around the world.
- CITI participated in the fall 2010 pNFS/NFSv4.1 interoperability workshop (Bake-a-thon), held in Hopkinton, MA.

5.8 University of California at Santa Cruz

In this last reporting period, UC Santa Cruz investigated several areas under the Petascale Data Storage Institute statement of work, primarily scalable metadata service and browsing and storage reliability and security. We also explored related areas such as long-term archival storage performance and access models for non-volatile memories.

Our research effort in providing scalable metadata services for petascale storage have been very successful to date, as measured both by publications and by research prototypes. We explored and evaluated different approaches to building scalable indexes that handle both file metadata searches and content-based searches. We are using a partitioned approach to metadata indexing that divides metadata from a large file system - we have conducted tests with file systems of more than a half billion files - into multiple partitions, each of which can be searched in parallel. In addition, the file system need not search partitions that cannot contain desired files; this determination is based on "summaries" of each partition that can be very quickly analyzed for each file system query. Our approach is 10-1000 times faster than existing database systems at metadata search, better allowing file system users to find locate specific files in a petabyte-scale storage system. In addition, our index requires far less space than traditional approaches built on standard databases, and is much more reliable, since failures in a portion of the index only require that portion to be rebuilt, avoiding a scan of the entire file system to rebuild a corrupt index. Our research was published at FAST 2009 and elsewhere. Recently, we submitted a follow-up paper to Eurosys 2010 describing our efforts to build a unified metadata and search index for a petascale file system, eliminating the need for a parallel database for metadata and removing the performance penalty for building an indexable file system.

A fundamental issue with scalable metadata handling is that its performance is limited by disk performance. We investigated the use of non-volatile RAM (NVRAM) technologies such as NAND flash, NOR flash, and phase-change RAM to alleviate this bottleneck. Our efforts to leverage NVRAM for high-performance metadata have been received positively. Our research on providing high levels of error detection and correction, necessary given the inherent bit-level unreliability of flash memory, was published at HotDep 2009 and EMSOFT 2009, and we intend to continue to explore issues with both flash memory and other non-volatile technologies such as phase-change RAM, transitioning to other funding sources as PDSI winds down.

Continuing our work on scalable metadata research, we built on our prior work from 2009 on index partitioning. During 2010, we looked at effective models for partitioning metadata search indexes, as well as creating an evaluation framework to describe the behavior of a given partitioning algorithm. We described a new model for effective file system partitioning which organizes files into partitions based on users and access rights, and evaluated it against a number of other popular algorithms for partitioning. We were able to demonstrate significant improvements over most existing partitioning algorithms, either in indexing time, search time, or both. Our work on security aware partitioning was published at MSST 2010.

We are also continuing to build on our work in securing petabyte-scale storage. We have already demonstrated an approach that provides strong authentication in a scalable parallel file system (Ceph), and have explored integration of this approach into the open-source Hadoop framework for map-reduce problems. We then turned our attention to encryption-based approaches to secure data in petascale storage, exploring techniques that can be used to secure data in a high-performance computing environment. We focused on the problem of mutual mistrust, in which HPC clients do not trust the storage and the storage system does not trust the HPC clients. In such an environment, our approach must guard against information leakage. To accomplish this goal, we built on our prior work in scalable authentication, which focused on reducing critical-I/O-path security communication with the metadata server. We added keys into the metadata and data transfer protocols to protect data in transit, protect data at rest on the disk, and yet use only symmetric encryption and decryption operations to read or write data. Modern hardware commonly provides support for symmetric cryptographic procedures, so cryptography is expected to lightly impact performance. Our approach limits the compromise within a distributed computing cluster to only the nodes which have an effectively full compromise in their running operating system. Plaintext data within our protocol can only be intercepted within active work on a corrupted node. We have a reference implementation in the Hadoop Distributed File System, which we found suffers computation overhead from cryptography in Java. We then implemented the protocol in the Ceph high-performance file system to demonstrate the impact on performance. To evaluate with the workload where this security model is most relevant, we ported Ceph to be an underlying file system within the Hadoop MapReduce framework, which has been noted in a USENIX ;login: article. The code for Ceph under Hadoop is presently in a patch under public review.

We have also continued our investigation into power-efficient archival storage. We constructed a discrete event simulator which we used to test the impact various data placement techniques had upon energy use in a highly-heterogeneous, archival "write-once" storage system. We found several interesting results, including identifying situations where utilizing more devices in the storage system may counter-intuitively save power, and noting that under very low read and write rates, data placement policies have minimal impact as power usage is dominated. This work however, highlighted the lack of relevant archival workload data available, and was the inspiration for our current project of obtaining and analyzing workload data from real-world archival systems, including the archive at PDSI member Los Alamos National Laboratory. This project is currently in progress and on track for a submission to Usenix 2011.

Our exploration of power-efficient storage systems is also including techniques for reducing power consumption in active storage systems. We worked with Lee Ward at Sandia National Laboratory to investigate approaches that group blocks onto different sets of disks, allowing the file system to power down some disks while leaving all necessary files available on powered-up disks. By leveraging machine learning algorithms for identification and grouping of related blocks, we hope future work will develop techniques that allow the storage system to work with only a subset of disks powered on, reducing overall storage system power consumption.

In order to alleviate the problems of current approaches to access Storage Class Memories (SCMs) and exploit the characteristics of various SCM devices without either limiting the design flexibility or introducing additional overhead, we are proposing the use of an object-based model for SCMs. The object-based model offloads the storage management layer from file system to the underlying hardware without sacrificing efficiency. Thus, it can be implemented on different types of SCM devices and hybrid storage systems, while the file system does not have to be changed as new technology comes out. Our initial implementation of object-based SCM is based on flash memory. Specifically, we explore three data placement policies enabled by an object-based storage model, including one that separates data and metadata, and another that further extracts access time from metadata and stores them in different segments to reduce the overall cleaning overhead when a log structure is used for storage. Using simulations, we showed that cleaning overhead can be reduced significantly by separating data, metadata, and access time especially under a read-intensive workload. In the future we hope to port our object-based model for flash memory into the Linux kernel, allowing us to further explore the benefits of the object-based model, such as extent-based allocation and object-level reliability. We also hope to work on approaches to exploit the use of object-based model on other SCMs. By isolating device-specific technology behind an object-based interface, we will allow users of HPC systems to write their applications once and leave device details to the device, in contrast to many current approaches that require changes in data layout and even application behavior based on specific device characteristics.

PDSI at UC Santa Cruz has been particularly effective at training graduate students in petascale file systems and I/O for high-performance computing. Four PDSI-funded students completed their PhDs: Deepavali Bhagwat, Kevin Greenan, Andrew Leung, and Mark Storer. In addition, we have graduated several master's students who worked on PDSI-funded research.

UCSC's PDSI research is documented online in several locations. We have a UCSC-specific PDSI web site at http://www.pdsi.ucsc.edu/, and a PDSI-wide web site at http://www.pdsi-scidac.org/. Additional information about research at UCSC in storage is available at http://www.ssrc.ucsc.edu/. As the PDSI grant winds down, we are moving the results of our research into the HPC community and transitioning personnel and research projects to other funding, when possible. Our metadata research is being moved to NSF HECURA funding and another Department of Energy project, and our archival storage research is being funded by NSF. We are also moving our results into the commercial realm by working with several companies through the NSF-chartered Center for Research in Intelligent Storage (http://www.cris.us/). In addition, we are talking with Whamcloud about transitioning our research into future versions of the Lustre file system.

6 Summary Metrics

The volume of PDSI supported personnel, publishing, and outreach is shown here:

	FY07	FY08	FY09	FY10	FY11	Total
Faculty/Staff	27	28	26	23		104
Students	28	32	33	22		115
Total Supported	55	60	59	45	0	219
Journal	5	4	5	1		15
Conf+Worksp	27	27	26	16	4	100
Other pubs	6	49	22	5	3	85
Total pubs	38	80	53	22	7	200
Talks	99	69	61	29		258
Workshops	1	2	2	2	1	8

7 Report References

[Abd-El-Malek-PDL08-106] File System Virtual Appliances: Third-party File System Implementations without the Pain. Michael Abd-El-Malek, Matthew Wachs, James Cipar, Gregory R. Ganger, Garth A. Gibson, Michael K. Reiter. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-106, May 2008.

[Abd-El-Malek-PDL-09-102] File system virtual appliances: Portable file system implementations. Michael Abd-El-Malek, Matthew Wachs, James Cipar, Karan Sanghi, Gregory R. Ganger, Garth A. Gibson, Michael K. Reiter. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-102, March 2009.

[CFDR] USENIX, The Computer Failure Data Respository (CFDR), http://cfdr.usenix.org.

[Dayal-08] S. Dayal, "Characterizing HEC Storage Systems at Rest," Carnegie Mellon University Parallel Data Laboratory CMU-PDL-08-109, 2008.

[Hase08-PDL08-107] "User Level Implementation of Scalable Directories (GIGA+)." Sanket Hase, Aditya Jayaraman, Vinay K. Perneti, Sundararaman Sridharan, Swapnil V. Patil, Milo Polte, Garth A. Gibson. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-107, May 2008.

[Hendricks-PDL06-104] "Eliminating Cross-server Operations in Scalable File Systems," James Hendricks, Shafeeq Sinnamohideen, Raja R. Sambasivan, Gregory R. Ganger. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-105, May 2006.

[LANL-FAILURE-07] Los Alamos National Laboratory, Operational Data to Support and Enable Computer Science Research, http://institutes.lanl.gov/data/fdata.

[Patil08-PDL08-110] "GIGA+ : Scalable Directories for Shared File Systems." Swapnil Patil, Garth Gibson. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-110. Oc-tober 2008.

[Phanishayee08] "Measurement and Analysis of TCP Throughput Collapse in Cluster-based Storage Systems," Amar Phanishayee, Elie Krevat, Vijay Vasudevan, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Srinivasan Seshan. 6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA. Supercedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-105, September 2007.

[Schroeder-07] B. Schroeder and G.A. Gibson, "Understanding disk failure rates: What does an MTTF of 1,000,000 hours mean to you?," ACM Transactions on Storage, 3(3):8, 2007.

[Schroeder-SciDAC07] Understanding Failures in Petascale Computers. Bianca Schroeder, Garth A. Gibson. SciDAC 2007. Journal of Physics: Conference Series 78 .

[Tantisiriroj-PDL08-114] "Data-intensive file systems for Internet services: A rose by any other name ...," Wittawat Tantisiriroj, Swapnil Patil, Garth Gibson. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-114. October 2008.

[Vasudevan09-PDL09-101] A (In)Cast of Thousands: Scaling Datacenter TCP to Kiloservers and Gigabits. Vijay Vasudevan, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-101, February 2009.

[Wachs08-PDL08-113] "Co-scheduling of Disk Head Time in Cluster-based Storage," Matthew Wachs, Gregory R. Ganger. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-113. October 2008.

[Wachs-FAST07] "Argon: Performance Insulation for Shared Storage Servers," Matthew Wachs, Michael Abd-El-Malek, Eno Thereska, Gregory R. Ganger. Proc. of the 5th USENIX Conf. on File and Storage Technologies (FAST'07), Feb. 2007, San Jose CA.

[WISH09] "Enabling Enterprise Solid State Disks Performance," Milo Polte, Jiri Simsa, Garth Gibson, Proc. of the First Workshop on Integrating Solid-State Memory into the Storage Hierarchy, held in conjunction with ASPLOS 2009, March 7, 2009, Washington DC.

8 Appendix 1: Publishing Details

FY07 Journals

Abd-El-Malek, Michael, William V. Courtright II, Chuck Cranor, Gregory R. Ganger, James Hendricks, Andrew J. Klosterman, Michael Mesnier, Manish Prasad, Brandon Salmon, Raja R. Sambasivan, Shafeeq Sinnamohideen, John D. Strunk, Eno Thereska, Matthew Wachs, Jay J. Wylie. Early Experiences on the Journey Towards Self-* Storage. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, September 2006.

Farber R. 2007. "Data Management, the Victorian era child of the 21st century", PNNL-SA-53343, Pacific Northwest National Laboratory, Richland, WA. published in Scientific Computing, vol. 24 no.4, March 2007

Farber R. 2007. "Finding the Insight to Balance Science-driven Computation and Experiment for Innovation" PNNL-SA-54125, Pacific Northwest National Laboratory, Richland, WA. published in Innovation: America's Journal of Technology Commercialization, vol. 5 no. 24, April/May 2007.

Schroeder, Bianca, Garth A. Gibson. Understanding Failures in Petascale Computers. SciDAC 2007. Journal of Physics: Conference Series 78 (2007) 012022, June 2007.

Zhang, Jiaying and Peter Honeyman, "A Replicated File System for Grid Computing," Concurrency and Computation: Practice and Experience (in press).

FY07 Conferences and Workshops

Greenan, Kevin, Ethan L. Miller, "PRIMS : Making NVRAM Suitable for Extremely Reliable Storage," short paper in Proceedings of the 3rd Workshop on Hot Topics in System Dependability (HotDep '07), June 2007, to appear.

Greenan, Kevin, Ethan L. Miller, "Reliability Mechanisms for File Systems Using Non-Volatile Memory as a Metadata Store," Proceedings of the 6th ACM & IEEE Conference on Embedded Software (EMSOFT '06), October 2006, pages 178-187.

Greenan, Kevin, Ethan L. Miller, Thomas Schwarz, Darrell D. E. Long, "Disaster Recovery Codes: Increasing Reliability with Large-Stripe Error Correction Codes," Proceedings of the 3rd International Workshop on Storage Security and Survivability (StorageSS 2007), held in conjunction with the 14th ACM Conference on Computer and Communications Security (CCS 2007), October 2007.

Hendricks, James, Gregory R. Ganger, Michael K. Reiter. Low-overhead Byzantine Fault-tolerant Storage. Proceedings of the Twenty-First ACM Symposium on Operating Systems Principles (SOSP 2007), Stevenson, WA, October 2007.

Hendricks, James, Gregory R. Ganger, Michael K. Reiter. Verifying Distributed Erasure-coded Data. Proceedings of the Twenty-Sixth Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2007), Portland, August 2007.

Hildebrand, Dean and Peter Honeyman, "Direct-pNFS: Scalable, transparent, and versatile access to parallel file systems," to appear in Proc. 16th IEEE International Symp. on High Performance Distributed Computing (HPDC 2007), Monterey. June 2007.

Hildebrand, Dean, Peter Honeyman, and W.A. (Andy) Adamson, "pNFS and Linux: Working towards a Heterogeneous Future," in Proc. 8th LCI International Conf. on High-Performance Clustered Computing, South Lake Tahoe. May 2007.

Kasick, Michael P., Priya Narasimhan, Kevin Atkinson, Jay Lepreau. Towards Fingerpointing in the Emulab Dynamic Distributed System. Proceedings of the 3rd USENIX Workshop on Real, Large Distributed Systems (WORLDS '06), Seattle, WA. Nov. 5, 2006.

Leung, Andrew, Eric Lalonde, Jacob Telleen, James Davis, Carlos Maltzahn, “Using Comprehensive Analysis for Performance Debugging in Distributed Storage Systems,” Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007), September 2007, pages 281-286.

Leung, Andrew, Ethan L. Miller, and Stephanie Jones, "Scalable Security for Petascale Parallel File Systems," submitted to SC '07, Reno, NV, November 2007.

Mesnier, Michael P., Matthew Wachs, Raja R. Sambasivan, Alice X. Zheng, Gregory R. Ganger. Modeling the Relative Fitness of Storage. SIGMETRICS'07, June 12-16, 2007, San Diego, California, USA.ACM. Awarded Best Paper.

Mesnier, Michael P., Matthew Wachs, Raja R. Sambasivan, Alice X. Zheng, Gregory R. Ganger. Modeling the Relative Fitness of Storage. SIGMETRICS’07, June 12–16, 2007, San Diego, California, USA.

Mesnier, Michael, Matthew Wachs, Raja R. Sambasivan, Julio Lopez, James Hendricks, Gregory R. Ganger. //TRACE: Parallel Trace Replay with Approximate Causal Events. Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), February 13-16, 2007, San Jose, CA. Supercedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-108, September 2006.

Pertet, Soila, Rajeev Gandhi and Priya Narasimhan. Fingerpointing Correlated Failures in Replicated Systems. USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML), Cambridge, MA. April 2007.

Pollack, Kristal, Darrell D. E. Long, Richard Golding, Ralph Becker-Szendy, Benjamin C. Reed, “Quota Enforcement for High-Performance Distributed Storage Systems,” Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST 2007), September 2007, pages 72-84.

Sambasivan, Raja R., Alice X. Zheng, Eno Thereska, Gregory R. Ganger. Categorizing and Differencing System Behaviours. Second Workshop on Hot Topics in Autonomic Computing. June 15, 2007. Jacksonville, FL.

Schroeder, Bianca, Garth A. Gibson. Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You? Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), February 13--16, 2007, San Jose, CA. Best Paper Award.
The above paper has also been featured in an article on slashdot, which so far has received more than 75,000 hits!

Schroeder, Bianca, Garth Gibson. A large scale study of failures in high-performance-computing systems. International Symposium on Dependable Systems and Networks (DSN 2006). One of the best DSN'06 papers invited to IEEE Transactions on Dependable and Secure Computing (TDSC).

Schroeder, Bianca, Garth Gibson. The computer failure data repository. Invited contribution to the Workshop on Reliability Analysis of System Failure Data (RAF'07) MSR Cambridge, UK, March 2007.

Shao, Minglong, Steven W. Schlosser, Stratos Papadomanolakis, Jiri Schindler, Anastassia Ailamaki, Gregory R. Ganger. MultiMap: Preserving Disk Locality for Multidimensional Datasets. IEEE 23rd International Conference on Data Engineering (ICDE 2007) Istanbul, Turkey, April 2007. Supercedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-102. March 2005.

Storer, Mark W., Kevin Greenan, Ethan L. Miller, Kaladhar Voruganti, “POTSHARDS: Secure Long-Term Storage Without Encryption,” Proceedings of the 2007 USENIX Technical Conference, June 2007, to appear.

Thereska, Eno, Dushyanth Narayanan, Anastassia Ailamaki, Gregory R. Ganger. Observer: Keeping System Models from Becoming Obsolete. Second Workshop on Hot Topics in Autonomic Computing. June 15, 2007. Jacksonville, FL. Supercedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-101, January 2007.

Wachs, Matthew, Michael Abd-El-Malek, Eno Thereska, Gregory R. Ganger. Argon: Performance Insulation for Shared Storage Servers. Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), February 13--16, 2007, San Jose, CA.

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

Weil, Sage, 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.

Zhang, Jiaying and Peter Honeyman, “Consistent Replication for Grid Computing,” in Proc. 4th International Workshop on Middleware for Grid Computing, Melbourne. November 2006.

Zhang, Jiaying and Peter Honeyman, “Hierarchical Replication Control in a Global File System,” in Proc. 7th IEEE International Symp. on Cluster Computing and the Grid (CCGrid07), Rio de Janeiro. May 2007.

FY07 Other

Baenziger, Clay, Bruce Bugbee, Ryan Ford, Charlie, Grammon. “LANL Supercomputing Data Analysis”, Colorado School of Mines Technical Report, 2007.

Davis, Olen, Kari Macklin, Baily Kelly “Parallel Search using multiple Google Desktops in Parallel”, Colorado School of Mines, Technical report, 2007.

Hildebrand, Dean “Distributed Access to Parallel File Systems,” Ph.D. dissertation, University of Michigan, Ann Arbor, February 2007.

Lalonde, Eric. “A Characterization of LANL HPC Systems”, Masters Thesis, University of California, Santa Cruz. 2007

Mehech, Max “The Impact of Failures on Large Distributed Storage Systems,” Technical Report UCSC-SSRC-07-10, August 2007.

Zhang, Jiaying “Network Transparency in Wide Area Collaborations,” Ph.D. dissertation, University of Michigan, Ann Arbor, May 2007.

FY08 Journals

Gibson, Garth , Bianca Schroeder, Joan Digney. “Failure Tolerance in Petascale Computers.” CTWatch Quarterly, vol. 3 no. 4. Volume on Software Enabling Technologies for Petascale Science. November 2007. www.ctwatch.org

Schroeder, Bianca , Garth A. Gibson. “Understanding Disk Failure Rates: What does an MTTF of 1,000,000 hours mean to you?” ACM Transactions on Storage (TOS), Volume 3 Issue 3, October 2007.

Storer, Mark W., Kevin Greenan, Ethan L. Miller, Kaladhar Voruganti. ”Pergamum: Energy-efficient Archival Storage with Disk Instead of Tape.” The USENIX Magazine 33(3), June 2008.

Zhang, Jiaying and Peter Honeyman, “A Replicated File System for Grid Computing,” Concurrency and Computation: Practice and Experience 20:9 (June 2008), pp. 1113–1130. DOI 10.1002/cpe.v20:9.

FY08 Conferences and Workshops

Bairavasundaram, L., G. Goodson, B. Schroeder, A. Arpaci-Dusseau, R. Arpaci-Dusseau, “An Analysis of Data Corruption in the Storage Stack.” 6th Usenix Conference on File and Storage Technologies (FAST 2008).

Curry, Matthew (University of Alabama at Birmingham, USA); Lee Ward (Sandia National Laboratories, USA); Tony Skjellum (University of Alabama Birmingham, USA); Ron Brightwell (Sandia National Laboratories, USA). “Accelerating Reed-Solomon Coding in RAID Systems with GPUs.” 22nd IEEE International Parallel and Distributed Processing Symposium, April 14-18, 2008, Miami, FL.

Curry, Matthew L., H. Lee Ward, Anthony Skjellum, and Ron Brightwell, University of Alabama at Birmingham and Sandia National Laboratory. Arbitrary Dimension Reed-Solomon Coding and Decoding for Extended RAID on GPUs. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Greenan, Kevin, Ethan L. Miller, Jay Wylie. “Reliability of XOR-based erasure codes on heterogeneous devices.” Proceedings of the 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2008), June 2008.

Greenan, Kevin, Ethan L. Miller, Thomas Schwarz, Darrell D. E. Long, “Disaster Recovery Codes: Increasing Reliability with Large-Stripe Error Correction Codes,” Proceedings of the 3rd International Workshop on Storage Security and Survivability (StorageSS 2007), held in conjunction with the 14th ACM Conference on Computer and Communications Security (CCS 2007), October 2007.

He, Yun (Helen), William T.C. Kramer, Jonathan Carter, Nicholas Cardo, Franklin: User Experiences, Proceedings of the Cray User Group 2008, Helsinki, Finland, May 5-8, 2008

Koren, Jonathan, Yi Zhang, Sasha Ames, Andrew Leung, Carlos Maltzahn, Ethan L. Miller, “Searching and Navigating Petabyte Scale File Systems Based on Facets,” Proceedings of the 2007 ACM Petascale Data Storage Workshop (PDSW 07), November 2007.

Koren, Jonathan, Yi Zhang, Xue Liu, “Personalized Interactive Faceted Search,” Proceedings of the 17th International Conference on the World Wide Web (WWW 2008), April 2008.

Kramer, William T.C. Yun (Helen) He, Jonathan Carter, Josephy Glenski, Lynn Rippe, Nicholas Cardo, Holistic Evaluation of Lightweight Operating Systems using the PERCU Method, LBNL Technical Report Number TBD, July 2008

Kramer, William, et al. Report of the Petascale Systems Integration Workshop, San Francisco, California, Published November 2007

Kramer, William, NERSC 2016—Extreme Computation and Data for Science, Proceedings of the Cray User Group 2008, Helsinki, Finland, May 5-8, 2008

Krevat, E., V. Vasudevan, A. Phanishayee, D. Andersen, G. Ganger, G. Gibson, S. Seshan. “On Application-level Approaches to Avoiding TCP Throughput Collapse in Cluster-Based Storage Systems.” Proceedings of the 2nd international Petascale Data Storage Workshop (PDSW ‘07) held in conjunction with Supercomputing '07. November 11, 2007, Reno, NV.

Leung, Andrew W. and Ethan L. Miller, University of California, Santa Cruz. Scalable Full-Text Search for Petascale File Systems. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Leung, Andrew, Ethan L. Miller, and Stephanie Jones, “Scalable Security for Petascale Parallel File Systems,” SC ‘07, Reno, NV, November 2007.

Leung, Andrew, Shankar Pasupathy, Garth Goodson, Ethan L. Miller, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings of the 2008 USENIX Technical Conference, June 2008.

Mackey, Grant, Saba Sehrish, John Bent, Jun Wang, University of Central Florida and Los Alamos National Laboratory. Introducing Map-Reduce to High End Computing. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008. Storer, Mark W., Kevin M. Greenan, Ian F. Adams, Ethan L. Miller, Darrell D. E. Long, Kaladhar Voruga, University of California, Santa Cruz. Logan: Automatic Management for Evolvable, Large-Scale, Archival Storage. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Oldfield, R.A., S. Arunagiri, P.J. Teller, S. Seelam, M.R. Varela, R. Riesen, P.C. Roth, “Modeling the Impact of Checkpoints on Next-Generation Systems,” IEEE Conference on Mass Storage Systems and Technologies, San Diego, California, Nov. 2007.

Patil, Swapnil V., Garth A. Gibson, Sam Lang, Milo Polte. “GIGA+: Scalable Directories for Shared File Systems.” Proceedings of the 2nd international Petascale Data Storage Workshop (PDSW '07) held in conjunction with Supercomputing '07. November 11, 2007, Reno, NV.

Phanishayee, Amar, Elie Krevat, Vijay Vasudevan, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Srinivasan Seshan. “Measurement and Analysis of TCP Throughput Collapse in Cluster-based Storage Systems.” 6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA.

Polte, Milo, Jiri Simsa, Wittawat Tantisiriroj, Garth Gibson. Fast Log-based Concurrent Writing of Checkpoints. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Roth, P.C. “Characterizing the I/O Behavior of Scientific Applications on the Cray XT,” 2007 Petascale Data Storage Workshop, co-located with SC07, Reno, Nevada, Nov. 2007.

Sambasivan, Raja R., Alice X. Zheng, Eno Thereska, Gregory R. Ganger. Categorizing and Differencing System Behaviours. Second Workshop on Hot Topics in Autonomic Computing. June 15, 2007. Jacksonville, FL.

Simsa, Jiri, Milo Polte, Garth Gibson. Comparing Performance of Solid State Devices and Mechanical Disks. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Storer, Mark W., Kevin Greenan, Ethan L. Miller, Kaladhar Voruganti, “Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage,” Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST '08), February 2008, pages 1-16.

Strunk, John D., Eno Thereska, Christos Faloutsos, Gregory R. Ganger. Using Utility to Provision Storage Systems. 6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA

Weil, Sage, Andrew Leung, Scott A. Brandt, Carlos Maltzahn, “RADOS: A Fast, Scalable, and Reliable Storage Service for Petabyte-scale Storage Clusters,” Proceedings of the ACM Petascale Data Storage Workshop 2007 (PDSW 07), November 2007.

Zhang, Jiaying and Peter Honeyman, “Performance and Availability Tradeoffs in Replicated File Systems,” to appear in Proc. International Workshop on Resiliency in High Performance Computing (RESILIENCE 2008), in conjunction with the 8th IEEE International Symposium on Cluster Computing and Grid (CCGRID 2008), Lyon (May 2008).

FY08 Other

Abd-El-Malek, Michael, Matthew Wachs, James Cipar, Gregory R. Ganger, Garth A. Gibson, Michael K. Reiter. File System Virtual Appliances: Third-party File System Implementations without the Pain. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-106, May 2008.

Ames, Sasha, Maya Gokhale, Carlos Maltzahn, Ethan L. Miller, UCSC. Queriable File Systems for Metadata Management. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Brown, DML, and GR Smith. 2008. "MPP2 Syslog Data (2006-2008)." PNNL-SA-61371 Pacific Northwest National Laboratory, Richland, WA.

Brown, DML. 2008. Final MPP2 Failure Data . PNNL-17833, Pacific Northwest National Laboratory, Richland, WA.

Cipar, Jim , Greg Ganger, Garth Gibson, Julio Lopez, Michael Stroucken, Wittawat Tantisiriroj, Dave O’Hallaron, Michael Kozuch, Michael Ryan, Steve Schlosser, Doug Cutting, Jay Kistler, Thomas Kwan. Tashi: Open-source Cloud Computing on Big Data. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Dayal, Shobhit, Garth Gibson, James Nunez, Evan J. Felix, Akbar Mokhtarani. “Filesystems Statistics Survey.” Carnegie Mellon University Parallel Data Lab Sping Industry Visit Day, May 15, 2008.

Dayal, Shobhit. Characterizing HEC Storage Systems at Rest. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-109, July 2008.

Farber, RM. 2008. “Back to the Future: The Return of Massively Parallel Systems.” PNNL-SA-59874, Pacific Northwest National Laboratory, Richland, WA.

Farber, RM. 2008. “Storage in Transition.” PNNL-SA-59313, Pacific Northwest National Laboratory, Richland, WA.

Felix, EJ. 2007. Statistical breakdown of MSCF production file systems in Oct 2007 . PNNL-17013, Pacific Northwest National Laboratory, Richland, WA.

Felix, EJ. 2007. “Statistical breakdown of MSCF production file systems in Oct 2007.” PNNL-17013, Pacific Northwest National Laboratory, Richland, WA.

Gibson, Garth, PDSI PIs. Petascale Data Management: Guided by Measurement. June 2008, Washington. D.C.

Gibson, Garth, PDSI. Petascale Data Management: Guided by Measurement. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Hase, Sanket, Aditya Jayaraman, Vinay K. Perneti, Sundararaman Sridharan, Swapnil V. Patil, Milo Polte, Garth A. Gibson. User Level Implementation of Scalable Directories (GIGA+). Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-107, May 2008.

Jain, Shailesh, Aditya Jayaraman, Sanket Hase, Sundar Sundaraman, Vinay Perneti, Swapnil Patil, Milo Polte, Garth Gibson. User-level Implementation of GIGA+ Scalable Directories. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Jayaraman, Aditya , Sanket Hase, Sundararaman Sridharan, Vinay K. Perneti, Swapnil Patil, Milo Polte, Garth Gibson. “User-level Implementation of GIGA+ Scalable Directories.” Carnegie Mellon University Parallel Data Lab Sping Industry Visit Day, May 15, 2008.

Jayaraman, Aditya, Sanket Hase, Sundararaman Sridharan, Vinay K. Perneti, Swapnil Patil, Milo Polte, Garth Gibson. User-level Implementation of GIGA+ Scalable Directories. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Klundt, Ruth, Marlow Weston, Lee Ward. I/O Tracing on Catamount. SAND2008-3684.

Leung, Andrew, Minglong Shao, Tim Bisson, Shankar Pasupathy, Ethan L. Miller, UC Santa Cruz. Spyglass: Metadata Search for Large-Scale Storage. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

López, Julio, Garth Gibson, Greg Ganger. DISC Experiences on the M45 Cluster. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Mesnier, Michael. On modeling the relative fitness of storage. PhD dissertation. Department of Electrical and Computer Engineering, Carnegie Mellon University. CMU Parallel Data Lab Technical Report CMU-PDL-07-108. December, 2007.

Mokhtarani, A., M. Andrews, W. Kramer, J. Hick, NERSC-LBNL. Large File System Backup, NGF Experience. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Mokhtarani, Akbar, Jason Hick, William T.C. Kramer, “Reliability Results of NERSC Systems”, LBN Report LBNL-430E

Mokhtarani, Akbar, Wayne Hurlbert, Nick Balthaser, Jason Hick, “Evaluation Report for a Tape Library Monitoring System,” LBNL report number pending.

Patil, Swapnil V. and Garth A. Gibson. “GIGA+: Scalable Directories for Shared File Systems (or, How to build directories with trillions of files).” Carnegie Mellon University Parallel Data Lab Sping Industry Visit Day, May 15, 2008.

Patil, Swapnil V. and Garth A. Gibson. GIGA+: Scalable Directories for Shared File Systems (or, How to build directories with trillions of files). 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Patil, Swapnil V. and Garth A. Gibson. GIGA+: Scalable Directories for Shared File Systems (or, How to build directories with trillions of files). 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Patil, Swapnil, Garth Gibson. GIGA+ : Scalable Directories for Shared File Systems. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-110. October 2008.

Patil, Swapnil, Garth Gibson. Large-scale Evaluation of GIGA+ Scalable Directories (using the FUSE user-level prototype). 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

PDSI PIs. Petascale Data Management: Guided by Measurement. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Petascale Data Storage Institute PIs. "High-Performance Metadata Indexing and Search in Petascale Data Storage Systems," SciDAC Conference, Seattle, WA, July 2008.

Petascale Data Storage Institute PIs. “PDSI Data Releases and Repositories.” Petascale Data Storage BoF Session at FAST '08. Feb 26-29, 2008, San Jose, CA.

Petascale Data Storage Institute PIs. “PDSI Shared Information Resources for HEC Storage.” Carnegie Mellon University Parallel Data Lab Sping Industry Visit Day, May 15, 2008 and ASCR PI meeting, March 31, 2008, Denver, CO.

Polte, Milo, Jiri Simsa, Wittawat Tantisiriroj, Garth Gibson. Log-structured Files for Fast Checkpointing. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Roth, P.C. and Jeffrey S. Vetter, “Understanding and Optimizing I/O Behavior of Scientific Applications: Petascale Data Storage Institute Activities at Oak Ridge National Laboratory” (poster), 2007 Petascale Data Storage Workshop, co-located with SC07, Reno, Nevada, Nov. 2007.

Simsa, Jiri , Milo Polte, Garth Gibson. Efficient Data Placement in a Hybrid Storage Architecture. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Simsa, Jiri, Garth Gibson, Randy Bryant. Formal Verification of Parallel File Systems. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Tantisiriroj, Wittawat, Garth Gibson. “Network File System (NFS) in High Performance Networks.” Carnegie Mellon University Parallel Data Lab Sping Industry Visit Day, May 15, 2008.

Tantisiriroj, Wittawat, Garth Gibson. Network File System (NFS) in High Performance Networks. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Tantisiriroj, Wittawat, Garth Gibson. Network File System (NFS) in High Performance Networks. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Tantisiriroj, Wittawat, Swapnil Patil, Garth Gibson. Crossing the Chasm: Sneaking a Parallel File System into Hadoop. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX. Nov. 17, 2008.

Tantisiriroj, Wittawat, Swapnil Patil, Garth Gibson. Crossing the Chasm: Sneaking a Parallel File System into Hadoop. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Tantisiriroj, Wittawat, Swapnil Patil, Garth Gibson. Data-intensive file systems for Internet services: A rose by any other name ... Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-114. October 2008.

UCSC Team. “Scalable Security for Petascale Parallel File Systems.” ASCR PI meeting, March 31, 2008,

Wachs, Matthew , Gregory R. Ganger. Co-scheduling of Disk Head Time in Cluster-based Storage. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-113. October 2008.

Wachs, Matthew, Elie Krevat, Mikhail Chainani, Chandramouli Rangarajan, Aditya Sethuraman, Greg Ganger. Performance Insulation in Shared Storage Clusters. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Wachs, Matthew, Greg Ganger. Co-scheduling of Disk Head Time in Cluster-based Storage. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Wachs, Matthew, Spencer Whitman, Deepti Chheda, Michael Abd-El-Malek, Eno Thereska, Greg Ganger. Argon: Performance Insulation for Shared Storage. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Yu, Weikuan, Jeffrey Vetter, Oak Ridge National Laboratory. Parallel I/O on the Cray XT. 3rd Petascale Data Storage Workshop held in conjunction with SC08, Austin, TX.

FY09 Journals

Grider, G., Nunez, J., Bent, J., Poole, S., Ross, R., and Felix, E. Coordinating government funding of file system and I/O research through the high end computing university research activity. SIGOPS Oper. Syst. Rev. 43, 1 (Jan. 2009), 2-7.

Leung, Andrew, Minglong Shao, Timothy Bisson, Shankar Pasupathy, Ethan L. Miller, Spyglass: Metadata Search for Large-Scale Storage Systems, ;login: - The USENIX Magazine 34(3), June 2009.

Molina-Estolano, E, C. Maltzahn, J. Bent and S. A. Brandt. Building a parallel file system simulator. Journal of Physics: Conference Series Volume 180 Number 1. 2009.

Storer, Mark W., Kevin Greenan, Ethan L. Miller, Kaladhar Voruganti, POTSHARDS - A Secure, Long-Term Storage System, ACM Transactions on Storage 5(2), June 2009.

Turner AS, KM Regimbal, MA Showalter, WA De Jong, CS Oehmen, ER Vorpagel, EJ Felix, RJ Rousseau, and TP Straatsma. 2009. EMSL Spawns Chinook and Looks to be Major Contributor in Supercluster Computing Community. PNNL-SA-64529, Pacific Northwest National Laboratory, Richland, WA. [Unpublished]

FY09 Conferences and Workshops

Adams, Ian, Darrell D. E. Long, Ethan L. Miller, Shankar Pasupathy, Mark W. Storer, Maximizing Efficiency By Trading Storage for Computation, Proceedings of the Workshop on Hot Topics in Cloud Computing (HotCloud '09), June 2009.

Bent, John, Garth Gibson, Gary Grider, Ben McClelland, Paul Nowoczynski, James Nunez, Milo Polte, Meghan Wingate. PLFS: A Checkpoint Filesystem for Parallel Applications. Supercomputing '09, November 15, 2009. Portland, Oregon.

Brandt, Scott A., Carlos Maltzahn, Neoklis Polyzotis, Wang-Chiew Tan, Fusing Data Management Services with File Systems. 4th Petascale Data Storage Workshop (PDSW 09), Portland, OR. November 15, 2009.

Buck, Joe, Noah Watkins, Carlos Maltzahn, Scott A. Brandt, Abstract Storage: Moving File Format- Specific Abstractions into Petabyte-Scale Storage Systems, 2nd International Workshop on Data- Aware Distributed Computing (in conjunction with HPDC-18), Munich, Germany, June 9, 2009.

Bung Chen, Hsing, Sarah Ellen Michalak, John Bent, and Gary Grider. FuseRBFS - A Fuse Based Ring-Buffer File System Design for Data Intensive Scientific Computing. In The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA09), Los Vegas, Nevada, July 2009.

Fan, Bin, Wittawat Tantisiriroj, Lin Xiao, Garth Gibson. DiskReduce: RAID for Data-Intensive Scalable Computing. 4th Petascale Data Storage Workshop held in conjunction with Supercomputing '09, November 15, 2009. Portland, Oregon. Supersedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-112, November 2009.

Gibson, Garth, Bin Fan, Swapnil Patil, Milo Polte, Wittawat Tantisiriroj, Lin Xiao. Understanding and Maturing the Data-Intensive Scalable Computing Storage Substrate. Microsoft Research eScience Workshop 2009, Pittsburgh, PA, October 16-17, 2009.

Greenan, Kevin, Darrell D. E. Long, Ethan L. Miller, Thomas Schwarz, Avani Wildani, Building Flexible, Fault-Tolerant Flash-based Storage Systems, Proceedings of the Fifth Workshop on Hot Topics in System Dependability (HotDep 2009), June 2009.

Jin, Keren, Ethan L. Miller, The Effectiveness of Deduplication on Virtual Machine Disk Images, Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, May 2009.

Kang, Yangwook, Ethan L. Miller, Adding Aggressive Error Correction to a High-Performance Flash File System, Proceedings of the 9th ACM/IEEE Conference on Embedded Software (EMSOFT '09), October 2009.

Kasick, Michael P., Keith A. Bare, Eugene E. Marinelli III, Jiaqi Tan, Rajeev Gandhi, Priya Narasimhan. System-Call Based Problem Diagnosis for PVFS. Proceedings of the 5th Workshop on Hot Topics in System Dependability (HotDep '09). Lisbon, Portugal. June 2009.

Leung, Andrew, Ethan L. Miller, Scalable Full-Text Search for Petascale File Systems, Proceedings of the 2008 Petascale Data Storage Workshop (PDSW 08), November 2008.

Leung, Andrew, Minglong Shao, Timothy Bisson, Shankar Pasupathy, Ethan L. Miller, Spyglass: Fast, Scalable Metadata Search for Large-Scale Storage Systems, Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST '09), February 2009.

Mitchell, Christopher, James Nunez, and Jun Wang. Overlapped Checkpointing with Hardware Assist. In Cluster 2009, New Orleans, Louisiana, August 2009.

Molina-Estolano, Esteban, Carlos Maltzahn, Scott Brandt, and John Bent, Comparing the Perfor- mance of Different Parallel Filesystem Placement Strategies, Work-In-Progress Session of the Con- ference on File and Storage Technology (FAST 2009), San Francisco, CA, February 24-27, 2009.

Molina-Estolano, Esteban, Maya Gokhale, Carlos Maltzahn, John May, John Bent, and Scott Brandt. Mixing Hadoop and HPC Workloads on Parallel Filesystems. In *Petascale Data Storage Workshop at SC09 (PDSW09)*, Portland, Oregon, November 2009.

Nithin Nakka, Alok Choudhary, Ruth Klundt, Marlow Weston, Lee Ward. Detailed analysis of I/O traces of large scale applications. Proceedings of the 16th annual IEEE International Conference on High Performance Computing (HiPC 2009), Kochi (Cochin), India, December 16-19, 2009.

Paris, Jehan-Francois, Ahmed Amer, Darrell D. E. Long, Using storage class memories to increase the reliability of two-dimensional RAID arrays, Technical Report UCSC-SSRC-09-04, April 2009.

Paris, Jehan-Francois, Ahmed Amer, Darrell D. E. Long, Using Storage Class Memories to Increase the Reliability of Two-Dimensional RAID Arrays, Proceedings of the 17th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2009), September 2009.

Park, Stan and Kai Shen. A Performance Evaluation of Scientific I/O Workloads on Flash-Based SSDs. IASDS 2009.

Patil, Swapnil, Garth A. Gibson, Gregory R. Ganger, Julio Lopez, Milo Polte, Wittawat Tantisiroj, and Lin Xiao. In Search of an API for Scalable File Systems: Under the table or above it? USENIX HotCloud Workshop 2009. June 2009, San Diego CA.

Polte, Milo, Jay Lofstead, John Bent, Garth Gibson, Scott A. Klasky, Qing Liu, Manish Parashar, Norbert Podhorszki, Karsten Schwan, Meghan Wingate, Matthew Wolf. ...And eat it too: High read performance in write-optimized HPC I/O middleware file formats. 4th Petascale Data Storage Workshop held in conjunction with Supercomputing '09, November 15, 2009. Portland, Oregon. Supersedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-111, November 2009.

Polte, Milo, Jiri Simsa, Garth Gibson. Enabling Enterprise Solid State Disks Performance. 1st Workshop on Integrating Solid-state Memory into the Storage Hierarchy, March 7, 2009, Washington DC.

Vasudevan, Vijay, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Brian Mueller. Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication. SIGCOMM'09, August 17–21, 2009, Barcelona, Spain.

Vasudevan, Vijay, Hiral Shah, Amar Phanishayee, Elie Krevat, David Andersen, Greg Ganger, Garth Gibson. Solving TCP Incast in Cluster Storage Systems. FAST 2009 Work in Progress Report. 7th USENIX Conference on File and Storage Technologies. Feb 24-27, 2009, San Francisco, CA.

Wildani, Avani, Thomas Schwarz, Ethan L. Miller, Darrell D. E. Long, Protecting Against Rare Event Failures in Archival Systems, Proceedings of the 17th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2009), September 2009.

FY09 Other

Abd-El-Malek, Michael, Matthew Wachs, James Cipar, Karan Sanghi, Gregory R. Ganger, Garth A. Gibson, Michael K. Reiter. File System Virtual Appliances: Portable File System Implementations. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-102. May 2009.

Abd-El-Malek, Michael. File System Virtual Appliances. Ph.D. Dissertation. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-109, August 2009.

Ames, Sasha, Carlos Maltzahn, Ethan L. Miller, Quasar: A Scalable Naming Language for Very Large File Collections, Technical Report UCSC-SSRC-08-04, October 2008.

Ames, Sasha, Maya Gokhale, Carlos Maltzahn, A Metadata-Rich File System, Technical Report UCSC-SOE-09-32, University of California at Santa Cruz, November 2009.

Brown DML. 2009. Application Strace Data for Jan 2009. PNNL-SA-64613 Pacific Northwest National Laboratory, Richland, WA.

Brown DML. 2009. Input Decks for Scientific Application Tracing data. PNNL-SA-64885 Pacific Northwest National Laboratory, Richland, WA.

Farber RM. 2008. Cloud Computing Scientific Computing, to be published November/December 2009,. PNNL-SA-63046.

Farber RM. 2008. People Make Petaflop Computing Possible Scientific Computing, November/December 2008. PNNL-SA-63046.

Gibson G, D Long, P Honeyman, G Grider, J Shalf, P Roth, EJ Felix, and L Ward. 2009. Petascale Data Storage Institute. PNNL-18366, Pacific Northwest National Laboratory, Richland, WA.

Gibson, Garth, Milo Polte. Directions for Shingled-Write and Two-Dimensional Magnetic Recording System Architectures: Synergies with Solid-State Disks. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-104. May 2009.

Greenan, Kevin, "Reliability and Power-Efficiency in Erasure-Coded Storage Systems," Technical Report UCSC-SSRC-09-08, December 2009.

Hendricks, James Vincent. Efficient Byzantine Fault Tolerance for Scalable Storage and Services. July 2009. [PhD thesis]

Klosterman, Andrew. Delayed instantiation bulk operations for management of distributed, object-based storage systems. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-108. August 2009. [PhD thesis]

Leung, Andrew, "Organizing, Indexing, and Searching Large-Scale File Systems," Technical Report UCSC-SSRC-09-09, December 2009.

Leung, Andrew, Ian Adams, Ethan L. Miller, "Magellan: A Searchable Metadata Architecture for Large-Scale File Systems," Technical Report UCSC-SSRC-09-07, November 2009.

Nithin Nakka, Alok Choudhary, Ruth Klundt, Marlow Weston, Lee Ward. Trace Data used to develop the paper "Detailed analysis of I/O traces of large scale applications" (HiPC 2009). http://www.cs.sandia.gov/Scalable_IO/SNL_Trace_Data/index.html

Patil, Swapnil, Garth Gibson. GIGA+: Scalable Directories for Shared File Systems. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-110. October 2008.

Skouson GB, DE Cowley, JF Carr, and DML Brown. 2009. Chinook Syslog data for failure analysis. PNNL-SA-66617 Pacific Northwest National Laboratory, Richland, WA.

Storer, Mark W. Secure, Energy-Efficient, Evolvable, Long-Term Archival Storage, Technical Report UCSC-SSRC-09-01, March 2009. [PhD thesis]

Tantisiriroj, Wittawat, Swapnil Patil, Garth Gibson. Data-intensive file systems for Internet services: A rose by any other name ... Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-114. October 2008.

Vasudevan, Vijay, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson. A (In)Cast of Thousands: Scaling Datacenter TCP to Kiloservers and Gigabits. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-09-101, Feb. 2009.

Wildani, Avani, Thomas Schwarz, Ethan L. Miller, Darrell D. E. Long, Protecting Against Rare Event Failures in Archival Systems, Technical Report UCSC-SSRC-09-03, April 2009. Preliminary version of a paper that appeared in MASCOTS 2009.

FY10 Journals

Maltzahn, Carlos, Esteban Molina-Estolano, Amandeep Khurana, Alex J. Nelson, Scott A. Brandt, and Sage Weil, "Ceph as a scalable alternative to the Hadoop Distributed File System," ;login: The USENIX Magazine, vol. 35, no. 4, pp. 38–49, August 2010.

FY10 Conferences and Workshops

Abe, Yoshihisa, Garth Gibson. pWalrus: Towards Better Integration of Parallel File Systems into Cloud Storage. Workshop on Interfaces and Abstractions for Scientific Data Storage (IASDS10), co-located with IEEE Int. Conference on Cluster Computing 2010 (Cluster10), Heraklion, Greece, September 2010.

Adams, Ian, Ethan L. Miller, Mark W. Storer, "Examining Energy Use in Heterogeneous Archival Storage Systems," Proceedings of the 18th Annual Meeting of the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2010), August 2010, pages 297-306.

Amer, Ahmed, Darrell D. E. Long, Ethan L. Miller, Jehan-François Pâris and Thomas J. E. Schwarz. "Design Issues for a Shingled Write Disk System," Proceedings of the Conference on Mass Storage Systems and Technologies, Incline Village, Nevada: IEEE, May 2010.

Bigelow, David, Scott Brandt, John Bent, HB Chen. Mahanaxar: Quality of Service Guarantees in High-Bandwidth, Real-Time Streaming Data Storage. MSST10, May 2010, Incline Village, Nevada.

Chaarawi, Sara, Jehan-François Pâris, Ahmed Amer, Thomas Schwarz and Darrell D. E. Long. "Using a Shared Storage Class Memory Device to Improve the Reliability of RAID Arrays," Proceedings of the 5th International Workshop on Petascale Data Storage (PDSW10), held in conjunction with SC2010, November 2010.

Chen, Y., X.-H. Sun, R. Thakur, H. Song and H. Jin, "Improving Parallel I/O Performance with Data Layout Awareness," 2010 IEEE International Conference on Cluster Computing (Cluster'10), Heraklion, Greece, September 2010.

Chen, Y., H. Zhu, H. Jin and X.-H. Sun, "Improving the Effectiveness of Context-based Prefetching with Multi-order Analysis," 3rd International Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2'10), San Diego, California, September 2010.

Corderi, Ignacio, Thomas Schwarz, Ahmed Amer, Darrell D. E. Long and Jehan-François Pâris. "Self-Adjusting Two-Failure Tolerant Disk Arrays," Proceedings of the 5th International Workshop on Petascale Data Storage (PDSW10), held in conjunction with SC2010, November 2010.

Jin, H., Y. Chen, and X.-H. Sun, "Optimizing HPC Fault-Tolerant Environment: An Analytical Approach," 39th International Conference on Parallel Processing (ICPP'10), San Diego, California, September 2010.

Kang, Yangwook, Jingpei Yang, Ethan L. Miller, "Efficient Storage Management for Object-based Flash Memory," Proceedings of the 18th Annual Meeting of the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2010), August 2010.

Pâris, Jehan-François, Thomas J. E. Schwarz, Ahmed Amer and Darrell Long. "Improving Disk Array Reliability Though Expedited Scrubbing," Proceedings of the Fifth IEEE International Conference on Networking, Architecture, and Storage (NAS 2010), Macau: IEEE, July 2010, pp. 119–125.

Parker-Wood, Aleatha, Christina Strong, Ethan L. Miller, Darrell D. E. Long, "Security Aware Partitioning for Efficient File System Search", 26th IEEE Symposium on Massive Storage Systems and Technologies: Research Track (MSST 2010), May 2010.

Patil, Swapnil, Garth Gibson. Scale and Concurrency of GIGA+: File System Directories with Millions of Files. Proceedings of the 9th USENIX Conference on File and Storage Technologies (FAST '11), San Jose CA, February 2011. Supersedes Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-10-110, Sept. 2010.

Sehrish, Saba, Grant Mackey, Jun Wang, John Bent. MRAP: A Novel MapReduce-based Framework to Support HPC Analytics Applications with Access Patterns. HPDC10, June 2010, Chicago, Illinios.

Vijayakumar, K., F. Mueller, X. Ma, and P.C. Roth, "Scalable I/O Tracing and Analysis," 2009 Petascale Data Storage Workshop, Portland, Oregon, November 2009.

Wildani, Avani, Ethan L. Miller, "Semantic Data Placement for Power Management in Archival Storage," Proceedings of the 5th International Workshop on Petascale Data Storage (PDSW10), held in conjunction with SC2010, November 2010.

FY10 Other

Abd-El-Malek, Michael, Matthew Wachs, James Cipar, Karan Sanghi, Gregory R. Ganger, Garth A. Gibson, Michael K. Reiter. File System Virtual Appliances: Portable File System Implementations. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-105, April 2010.

Marti Bancroft, John Bent, Evan Felix, Gary Grider, James Nunez, Steve Poole, Robert Ross, Ellen Salmon, Lee Ward. Report on the High End Computing Interagency Working Group. (HECIWG) Sponsored File Systems and I/O Workshop HEC FSIO 2010.

Sambasivan, Raja R., Alice X. Zheng, Elie Krevat, Spencer Whitman, Gregory R. Ganger. Diagnosing Performance Problems by Visualizing and Comparing System Behaviours. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-103, February 2010.

Taborda, Ricardo, Julio López, Haydar Karaoglu, John Urbanic, Jacobo Bielak. Speeding Up Finite Element Wave Propagation for Large-Scale Earthquake Simulations. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-109, October 2010.

Zheng, Alice X., Elie Krevat, Spencer Whitman, Michael Stroucken, William Wang, Lianghong Xu, Gregory R. Ganger. Diagnosing Performance Changes by Comparing System Behaviours. Raja R. Sambasivan, Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-107. July 2010. Supersedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-103.

FY11 Journals

FY11 Conferences and Workshops

Lofstead, Jay, Milo Polte, Garth Gibson, Scott A. Klasky, Karsten Schwan, Ron Oldfield, Matthew Wolf, Qing Liu. Six Degrees of Scientific Data: Reading Patterns for Extreme Scale Science IO. 20th ACM Int. Symp. On High-Performance Parallel and Distributed Computing (HPDC'11), June 2011.

Patil, Swapnil, Garth Gibson.Scale and Concurrency of GIGA+: File System Directories with Millions of Files. Proceedings of the 9th USENIX Conference on File and Storage Technologies (FAST '11), San Jose CA, February 2011. Supersedes Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-10-110, Sept. 2010.

Sambasivan, Raja R., Alice X. Zheng, Michael De Rosa, Elie Krevat, Spencer Whitman, Michael Stroucken, William Wang, Lianghong Xu, Gregory R. Ganger. Diagnosing Performance Changes by Comparing Request Flows. 8th USENIX Symposium on Networked Systems Design and Implementation (NSDI'11). March 30 - April 1, 2011. Boston, MA.

Tantisiriroj, Wittawat, Swapnil Patil, Garth Gibson, Seung Woo Son, Samuel J. Lang, Robert B. Ross. On the Duality of Data-intensive File System Design: Reconciling HDFS and PVFS. SC11, November 12-18, 2011, Seattle, Washington USA. Supersedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-108. April 2011.

FY11 Other

Fan, Bin, Wittawat Tantisiriroj, Lin Xiao, Garth Gibson. DiskReduce: Replication as a Prelude to Erasure Coding in Data-Intensive Scalable Computing. Carnegie Mellon Univsersity Parallel Data Laboratory Technical Report CMU-PDL-11-112, October, 2011.

Gibson, Garth, Greg Ganger. Principles of Operation for Shingled Disk Devices. Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-107. April 2011.

Sambasivan, Raja R., Gregory R. Ganger. Automation Without Predictability is a Recipe for Failure. Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-11-101, January 2011.

9 Appendix 2: External Presentations and Outreach

FY07

Bent, John (HPC-5) “Devising Fault Tolerant I/O Models for Next Generation Super Computers,”.

Bent, John Fall 2006 Topics in Computer Science Seminar Series. Local (LANL Staff) and remote (UCSC faculty and graduate students) audience participation via Polycom transmission.

Bigelow, D., S. Iyer, T. Kaldewey, R. Pineiro, A. Povzner, S. Brandt, R. Golding (presenter), T. Wong, C. Maltzahn, Univ. of California, Santa Cruz, IBM-Almaden. “End-to-end Performance Management for Scalable Distributed Storage.” Petascale Data Storage Workshop Supercomputing ‘07, Reno, Nevada, Nov. 2007.

Brandt, Scott. “Ceph: A Scalable, High-Performance Distributed File System,” University of Rhode Island, Kingston, RI, September 2006.

Brandt, Scott. “Ceph: Scalable, Reliable, Secure, High-Performance Storage,” Aster Data Systems, Redwood Shores, CA, March 2007.

Brandt, Scott. “Ceph: Scalable, Reliable, Secure, High-Performance Storage,” University of California, San Diego, March 2007.

Brandt, Scott. “Ceph: Scalable, Reliable, Secure, High-Performance Storage,” University of New Mexico, Albuquerque, NM, January 2007.

Brown, David, PNNL. “Debian Lustre and PVFS Repository.” Short announcement at PDSW, co-located with SC07, Reno, Nevada, Nov, 2007.

Chen, Hsing-bung (HPC-5) “Scalable Server I/O Networking Architectures for Very Large-Scale Linux Clusters – PaScal I/O vs. Federated I/O,”

Cook, Danny (HPC-3) “Archival Storage at LANL: Past, Present and Future,”

Daly, John T. (HPC-4) “Failure Analysis from the Application's Point of View”,.

Daly, John T. Performance Challenges for Extreme Scale Computing. Los Alamos National Lab. SDI Seminar, Carnegie Mellon University. October 2007.

DeBardeleben, Nathan (HPC-4) “Dynamically Probing the Linux Kernel,”

Felix EJ, and DE Cowley. 2006. “PNNL EMSL Lustre Activities November 2006 .” Presented by Evan J. Felix (Invited Speaker) at Supercomputing 2006 - Lustre BOF, Tampa, FL on November 14, 2006. PNNL-SA-52864.

Felix EJ, and DE Cowley. 2006. “PNNL EMSL Lustre Activities November 2006 .” Presented by Evan Felix (Invited Speaker) at SciDAC Petascale data storage workshop, Tampa, FL on November 17, 2006. PNNL-SA-52864.

Felix EJ, and J Nieplocha. 2006. “Advanced Data Processing with Active Storage.” Presented by Evan Felix at Supercomputing 2006, Tampa, FL on November 14, 2006. PNNL-SA-52574.

Felix, E.J. “PNNL – Petascale Data Storage Institute Data release update FAST08.” 6th USENIX Conference on File and Storage Technologies, San Jose, CA on February 28, 2008. PNNL-SA-59485.

Felix, E.J. (presenter) and R. Farber. “Balancing Storage Bandwidth to Petaflops.” Supercomputing 2007, Reno, NV on November 10, 2007. PNNL-SA-57749.

Felix, EJ, and R Farber. 2007. "Balancing Storage Bandwidth to Petaflops." Presented by N/A at Supercomputing 2007, Reno, NV on November 10, 2007. PNNL-SA-57749.

Felix, Evan, PNNL. “fsstats Data Release.” Short announcement at PDSW, co-located with SC’07, Reno, Nevada, Nov, 2007.

Gibson, Garth. “An Opinion for the File I/O Panel,” NSA Advanced Computing Systems Workshop, June 2007, Washington DC.

Gibson, Garth. “Managing the Information Tsunami,” Director if National Intelligence (DNI) Open Source Conference, July 2007, Washington DC.

Gibson, Garth. “Parallel NFS,” Fifth Intelligent Storage Workshop, University of Minnesota, May 2007, Minneapolis MN.

Gibson, Garth. “Parallel NFS,” Sixth HLRS Workshop on Scalable Global Parallel File Systems, University of Stuttgart, April 2007, Stuttgart, Germany.

Gibson, Garth. “Petascale Data Storage Institute,” HECIWG FSIO 2007 Workshop, August 2007, Arlington VA.

Gibson, Garth. “Reflections on Failure in Post-Terascale Parallel Computing,” 2007 International Conference on Parallel Processing, September, 2007, XiAn, China.

Gibson, Garth. “Storage in 2031,” Storage in 2031 Seagate Workshop, June 2007, Pittsburgh PA.

Gibson, Garth. “Storage Trends., Workshop on Trends in Computing Performance, Computer Science and Telecommunications Board Committee on Sustaining Growth in Computing Performance, The National Academies, September 2007, Mountain View, CA.

Gibson, Garth. “Trends in Cluster Fault Tolerance and Parallel File System Standards,” Workshop on High Performance Computing in Geophysics, SEG2007, San Antonio TX.

Gibson, Garth. “Understanding Failure at Petascale+,” DARPA Exascale Computing Study, Storage and Applications, September 2007, Berkeley, CA.

Gibson, Garth. “Understanding Failure in Petascale Computers,” 2007 SciDAC Conference, June, 2007, Boston MA.

Grider, Gary. “HPC Storage, File Systems, and I/O past, present, and future,” Colorado School of Mines, Golden Co.

Hendricks, James. “Low-overhead Byzantine Fault-tolerant Storage.” 21st ACM Symposium on Operating Systems Principles (SOSP 2007), in Stevenson, WA.

Hendricks, James. “Verifying Distributed Erasure-coded Data.” 26th Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2007), Portland, OR.

Hildebrand, Dean. “Distributed, Scalable, and Independent Access to Parallel File Systems,” Argonne National Laboratory, Argonne, IL (October 2006).

Hildebrand, Dean. “pNFS and Linux: Working towards a Heterogeneous Future,” presented at 8th LCI International Conf. on High-Performance Clustered Computing, South Lake Tahoe (May 2007).

Honeyman, Peter, “Consistent Replication for Grid Computing,” presented at 4th International Workshop on Middleware for Grid Computing, Melbourne (November 2006).

Honeyman, Peter, “Hierarchical Replication Control in a Global File System,” presented at 7th IEEE International Symp. on Cluster Computing and the Grid (CCGrid07), Rio de Janeiro (May 2007).

Honeyman, Peter, “NFSv4 and Cluster File Systems,” Tutorial, FAST 07

Honeyman, Peter, “NFSv4 for Grid Computing,” IBM Haifa Research Lab (May 2007).

Honeyman, Peter, Graduate seminar: “Storage” (EECS 598), Winter 2007 semester.

Konwinski, Andrew. “A Universal Taxonomy for Categorizing Tracing Frameworks.” PDSI Workshop at Supercomputing 07 in Reno, NV.

Konwinski, Andrew. “A Universal Taxonomy for Categorizing Tracing Frameworks.” Petascale Data Storage Workshop, Supercomputing '07, Reno, Nevada, Nov. 2007.

Kramer, William T.C, Large System Evaluation at NERSC Evaluating Petascale Infrastructure Systems: Benchmarks, Models and Applications, panel Speaker on System Evaluation at SC 06, November, 2006 in Tampa, FL.

Kramer, William T.C, Organized and Chaired the Exotic Storage Technologies Panel session at SC 06, November 2006 Tampa FL.

Kramer, William T.C., Large Scale System Evaluation at NERSC, Workshop on Procurement Best Practices at SC 06, November, 2006 in Tampa, FL.

Kramer, William T.C., NERSC Experience and Plans for Petascale Data, the Petascale Data Storage Workshop at SC 06, November 18, 2006 in Tampa, FL.

Kramer, William T.C., NERSC Experiences: Implementation of a Facility Wide Global File System, presentation at 12th ECMWF (European Center for Mid-range Metrological Forecasting) Workshop on Use of High Performance Computing in Meteorology, - Reading England, October 30-November3 2006.

Kramer, William. NERSC. “Reliability Results of NERSC Systems.” Short announcement at PDSW, co-located with SC07, Reno, Nevada, Nov, 2007.

Mesnier, Mike. “Modeling the Relative Fitness of Storage.” SIGMETRICS’07, June 12–16, 2007, San Diego, California. Received the best paper award

Miller, Ethan. "Ceph: A Scalable, High-Performance Distributed File System," National Technology Alliance Data Retention and Processing Workshop, Herndon, VA, July 2007.

Miller, Ethan. "Disaster Tolerance Codes," Agami Systems, Sunnyvale, CA, September 2007.

Miller, Ethan. "POTSHARDS: Secure Long-Term Archival Storage Without Encryption," VMware, Palo Alto, CA, September 2007.

Miller, Ethan. "Reliability Mechanisms for File Systems Using Non-Volatile Memory as a Metadata Store," Network Appliance, Sunnyvale, CA, September 2007.

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption,” University of New Mexico, Albuquerque, NM, April 2007.

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption,” Winter 2007 SNIA Storage Security Workshop, San Diego, CA, Janaury 2007

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption,” IBM Almaden Research Lab, San Jose, CA, January 2007.

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption,” Data Domain, Santa Clara, CA, December 2006.

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption,” Brown University, Providence, RI, September 2006.

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption.” University of California, Santa Barbara, November 2007.

Miller, Ethan. “POTSHARDS: Secure Long-Term Archival Storage Without Encryption.” TRUST seminar, University of California, Berkeley, November 2007.

Miller, Ethan. “Recent Research at the Storage Systems Research Center.” Data Domain, Sunnyvale, CA, December 2007.

Miller, Ethan. “Scalable Security for Petascale Parallel File Systems.” SC07, Reno, NV, November 2007.

Miller, Ethan. “Using Massive Arrays of Idle Disks (MAIDs) for Long-Term Storage in Digital Libraries,” Digital Library Forum, Pasadena, CA, April 2007.

Miller, Ethan. Improving File System Performance and Reliability with Non-Volatile Memory,” eBay, San Jose, CA, February 2007.

Miller, Ethan. Visited the Data Storage Institute, affiliated with the National University of Singapore, in July 2007, gave five talks and discussed DSI-UCSC collaborations with DSI researchers.

- "Ceph: A Scalable, High-Performance Distributed File System"
- "Maat: Scalable Security for High Performance, Petascale Storage"
- "POTSHARDS: Secure Long-Term Archival Storage Without Encryption"
- "Store, Forget, and Check: Using Algebraic Signatures to Check Remotely Administered Storage"

Mokhtarani, Akbar. *NERSC System Reliability data*, HECIWG FSIO 2007 Workshop, August 2007, Arlington VA.

Nieplocha, J, J Piernas Canovas, and EJ Felix. 2007. "User-Space Implementation of Active Storage for Lustre Parallel Filesystem." Presented by Jarek Nieplocha at HP-CAST-9 Meeting (part of SC'07), Reno, NV on November 9, 2007. PNNL-SA-58043.

Nieplocha. J (presenter), J Piernas Canovas, and EJ Felix. “User-Space Implementation of Active Storage for Lustre Parallel Filesystem.” HP-CAST-9 Meeting (part of SC’07), Reno, NV on November 9, 2007. PNNL-SA-58043.

Nunez, James. “Data release update, including fsstats data and trace availability.” Petascale Data Storage Workshop, Supercomputing '07, Reno, Nevada, Nov. 2007.

Nunez, James. Data release update, including planned fsstats data and trace availability, HECIWG FSIO 2007 Workshop, August 2007, Arlington VA.

Nunez, James. Standards update talk given at SC06 for HECEWG POSIX Extensions

Patil, Swapnil V. (presenter), Garth A. Gibson, Sam Lang, Milo Polte. “GIGA+: Scalable Directories for Shared File Systems.” Petascale Data Storage Workshop, Supercomputing '07, Reno, Nevada, Nov. 2007.

Roth, P.C., “Characterizing the I/O Behavior of Scientific Applications on the Cray XT,” 2007 Petascale Data Storage Workshop, co-located with SC07, Reno, Nevada, Nov. 2007.

Sambasivan, Raja R.. “Categorizing and Differencing System Behaviours.” Second Workshop on Hot Topics in Autonomic Computing. June 15, 2007. Jacksonville, FL.

Schroeder, Bianca, U. of Toronto & CMU. “Computer Failure Data Repository.” Short announcement at PDSW, co-located with SC07, Reno, Nevada, Nov, 2007.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, Boston University.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, Brown University.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, University of Toronto.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, University of Waterloo.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, Purdue University.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, University of California, Riverside.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” April 2007, Cornell University.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” April 2007, Georgie Institute of Technology.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” April 2007, University of California, Berkeley.

Schroeder, Bianca. “The computer failure data repository (CFDR)”. HEC-IWG File Systems and I/O R&D Workshop, August 2007, Washington D.C.

Schroeder, Bianca. “Failures in the real world, ” Invited talk for SNIA (Storage Networking Industry Association) Developers Conference, September 2007, San Jose, CA.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” March 2007, Ohio State University.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” April 2007, Rice University.

Schroeder, Bianca. “From Web Servers to Databases to Storage systems: A Methodological Approach to System Design,” April 2007, Hewlett Packard Laboratories, Palo Alto, CA. Host: John Wilkes.

Storer, Mark. "POTSHARDS: Secure Long-Term Archival Storage Without Encryption," Sun Microsystems, Menlo Park, CA, September 2007.

Thereska, Eno. “Observer: Keeping System Models from Becoming Obsolete.” Second Workshop on Hot Topics in Autonomic Computing. June 15, 2007. Jacksonville, FL.

Ward, Lee. “GPU as RAID accelerator,” HECIWG FSIO 2007 Workshop, August 2007, Arlington VA.

Zhang, Jiaying. “Naming and Replication in Wide Area Collaborations”, EMC, Boston (March 2007).

Zhang, Jiaying. “Network Transparency in Wide Area Collaborations”, VMware, Palo Alto (December 2006).

FY08

Abd-El-Malek, Michael. File System Virtual Appliances. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Bent, John. “HEC FSIO and LANL FSIO.” 25th IEEE Symposium on Massive Storage Systems and Technologies.

Bent, John. Data release update, including fsstats data and trace availability. 2008 HECIWG FSIO Workshop in Arlington, VA.

Brandt, Scott, Maltzahn, Carlos. “Ceph: A Scalable, High-Performance Distributed File System.” University of Paderborn, Paderborn, Germany, February 2008.

Brandt, Scott, Maltzahn, Carlos. “Scalable Security in Ceph.” Winter 2008 SNIA Security Summit, San Jose, CA, January 2008.

Dayal, Shobhit. “File System Statistics.” Shobhit Dayal, CMU, Garth Gibson, CMU, Marc Unangst, Panasas. PDS BoF held in conjunction with USENIX FAST’08, San Jose, CA.

Felix, EJ. 2008. "PNNL – Petascale Data Storage Institute Data release update FAST08." Presented by Evan Felix (Invited Speaker) at 6th USENIX Conference on File and Storage Technologies, San Jose, CA on February 28, 2008. PNNL-SA-59485.

Felix, EJ. 2008. "PNNL – Petascale Data Storage Institute fsstats - Data release update." Presented by Evan Felix at HEC FSIO Workshop 2008, Arlington, VA on August 4, 2008. PNNL-SA-61771.

Ganger, Greg. Performance Insulation and Predictability for Shared Cluster Storage. HEC FSIO R&D Conference/HECURA FSIO PI Meeting '08, Arlington, VA. Aug 3 - Aug 6, 2008.

Gibson, Garth. “PDSI FAST 2008 BOF Introduction.” PDS BoF held in conjunction with USENIX FAST’08, San Jose, CA.

Gibson, Garth. Exascale and Clouds on the Horizon. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Gibson, Garth. Failure in Supercomputers and Supercomputer Storage. NSF/DOE Expedition Workshop/Toward Scalable Data Management. June 10, 2008. Washington, D.C.

Gibson, Garth. GIGA+: Scalable Directories for Shared File Systems. HEC FSIO R&D Conference/HECURA FSIO PI Meeting '08, Arlington, VA. Aug 3 - Aug 6, 2008.

Gibson, Garth. SciDAC PDSI Update. HEC FSIO R&D Conference/HECURA FSIO PI Meeting '08, Arlington, VA. Aug 3 - Aug 6, 2008.

Grider, Gary. “HEC FSIO Efforts.” UC Santa Cruz’ Storage Systems Research Center 6th Annual Research Review Meeting.

Hendricks, James. Byzantine Fault Tolerance for Storage and Services. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Kasick, Mike. Diagnosing Performance Problems in PVFS. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Klosterman, Andy. Implementation Strategies for Bulk Operations in Cluster-based Storage. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Kramer, William T.C., Challenges of Petascale System Integration into Large Scale Facilities, Petascale Systems Integration into Large Scale Facilities Workshop, San Francisco, CA, May 2007.

Kramer, William T.C., Cray’s XT4 Integration and Progress at NERSC, Cray Technical Workshop, February 26-27, 2008, San Francisco, CA

Kramer, William T.C., The New HPC Center – Where the Grid and Petascale Meet, 2007 Australian Partnership on Advanced Computing (APAC 07) Conference, (http://www.apac.edu.au/apac07/), October 8-12, 2007, Perth, Australia

Kramer, William, T.C., The Biggest Challenges for Petascale Systems, LCI Cluster Computing Workshop, April 29-May 1, 2008, Champaign, Illinois

Kramer, William. Cray's XT4 Integration and Progress at NERSC, Cray Technical Workshop, February 26-27, 2008, San Francisco, CA

Kramer, William. NERSC - Extreme Storage and Computation for Science, Keynote presentation at the Storage Network World Conference, April 7-10, 2008, Orlando FL

Kramer, William. NERSC. "Extreme Storage and Computation for Science." Keynote presentation at the Storage Network World Conference, April 7-10, 2008, Orlando FL

Krevat, E. (presenter), V. Vasudevan, A. Phanishayee, D. Andersen, G. Ganger, G. Gibson, S. Seshan. "On Application-level Approaches to Avoiding TCP Throughput Collapse in Cluster-Based Storage Systems." Petascale Data Storage Workshop Supercomputing '07, Reno, Nevada, Nov. 2007.

Krevat, Elie. "Measurement and Analysis of TCP Throughput Collapse in Cluster-based Storage Systems." Amar Phanishayee, Elie Krevat, Vijay Vasudevan, David G. Andersen, Gregory R. Ganger, Gibson, Garth A., Srinivasan Seshan. 6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA.

López, Julio. Early Experiences with Disc-style Computation. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Miller, Ethan. Panel member, "SSD in the Enterprise," Flash Memory Summit, Santa Clara, CA.

Miller, Ethan. "Pergamum: Building Evolvable, Reliable, Searchable, Energy-Efficient Petascale Archival Storage from Disks," Yahoo!, Santa Clara, CA, September 2008.

Miller, Ethan. "Pergamum: Evolvable Energy Efficient, Reliable, Disk-Based Archival Storage," invited talk at the 5th IEEE International Workshop on Storage Network Architecture and Parallel I/O, Baltimore, MD, September 2008.

Miller, Ethan. "Pergamum: Evolvable Energy Efficient, Reliable, Disk-Based Archival Storage," IBM off- site research review, San Jose, CA, June 2008.

Miller, Ethan. "Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage," LSI Logic, Boulder, CO, June 2008.

Miller, Ethan. "Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage," 8th Intelligent Storage Workshop, University of Minnesota, Minneapolis, MN, May 2008.

Miller, Ethan. "Ceph: An Open-Source Petabyte-Scale File System." Storage Systems Research Center, UCSanta Cruz. PDSI BoF at USENIX FAST08 in San Jose, CA

Miller, Ethan. "Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage." Symantec, Mountain View, CA, March 2008.

Miller, Ethan. "Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage." FAST 2008, San Jose, CA, February 2008.

Miller, Ethan. "Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage." NetApp, Sunnyvale, CA, February 2008.

Miller, Ethan. "Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage." University of Paderborn, Paderborn, Germany, February 2008.

Mokhtarani, Akbar. Large File System Backup, NERSC Global File System Experience, Poster session, HEC FSIO, Aug. 2008

Mokhtarani, Akbar. PDSI work at NERSC, presented at BoF at 2008 USENIX Conference, San Jose, Ca. February 2008.

Narasimhan, Priya. Towards Automated Problem Analysis of Large-Scale Storage Systems. HEC FSIO R&D Conference/HECURA FSIO PI Meeting '08, Arlington, VA. Aug 3 - Aug 6, 2008.

Nunez, James. “Data release update, including fsstats data and trace availability.” PDS BoF held in conjunction with USENIX FAST’08, San Jose, CA.

Nunez, James. “High End Computing File Systems and I/O (HEC FSIO): Coordinating the US Government Research Investments.” SIAM Conference on Parallel Processing for Scientific Computing.

Nunez, James. “High End Computing File Systems and I/O (HEC FSIO): Coordinating the US Government Research Investments. ”Oakridge National Lab’s I/O Workshop.

Nunez, James. “High End Computing File Systems and I/O (HEC FSIO): Coordinating the US Government Research Investments.” Texas A&M’s First Annual Texas Research Exchange Festival.

Nunez, James. “High End Computing File Systems and I/O (HEC FSIO): Coordinating the US Government Research Investments.” SIAM Conference on Parallel Processing for Scientific Computing.

Nunez, James. “LANL Data Release and I/O Forwarding Scalable Layer (IOFSL): Background and Plans.” talk given by James Nunez at the 2008 Intelligent Storage Workshop at the University of Minnesota.

Nunez, James. “Open Problems and Gaps in High End Computing File Systems and I/O (HEC FSIO).” Office of Science Computer Science Principal Investigator Meeting.

Nunez, James. “Open Problems and Gaps in High End Computing File Systems and I/O (HEC FSIO).” Office of Science Computer Science Principal Investigator Meeting.

Nunez, James. Data release update, including fsstats data and trace availability. PDSI Workshop at Supercomputing 07 in Reno, NV.

Nunez, James. Data release update, including fsstats data and trace availability. PDSI BoF at USENIX FAST08 in San Jose, CA

Patil, Swapnil. GIGA+: Scalable Directories for Shared File Systems. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Patil, Swapnil. Invited talk on “GIGA+: Scalable Directories for Shared File Systems” at the Conference on Scalability 2008 organized by Google in Seattle WA.

Peter Honeyman, Performance and Availability Tradeoffs in Replicated File Systems. Presented at RESILIENCE 2008. (Workshop on Resiliency in High Performance Computing, in conjunction with CCGRID), Lyon (May 2008).

Peter Honeyman, Storage Research at CITI, Poster Session, Open House for Prospective PhD Students on CSE Research, University of Michigan.

Polte, Milo. Log-structured Files for Fast Checkpointing. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008. Nov 3-5, 2008.

Sambasivan, Raja. Performance Diagnosis in Distributed Storage Systems. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Schmidt, KP, KM Fox, and EJ Felix. "Lustre Backup Filesystem." PNNL-SA-61642 Pacific Northwest National Laboratory, Richland, WA.

Schroeder, Bianca. “The Computer Failure Data Repository (CFDR).” PDS BoF held in conjunction with USENIX FAST’08, San Jose, CA.

Shalf, John. National Energy Research Scientific Computing Center (NERSC), LBNL. “I/O Requirements for HPC Applications: A User Perspective.” Special Presentation at PDS BoF held in conjunction with USENIX FAST’08, San Jose, CA.

Simsa, Jiri. Flash for High-end Storage. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Sinnamohideen, Shafeeq. Dynamic Scaling of Metadata Services. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Strunk, John D. “Using Utility to Provision Storage Systems.” John D. Strunk, Eno Thereska, Christos Faloutsos, Gregory R. Ganger. 6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA.

Tantisiriroj, Wittawat. Crossing the Chasm: Sneaking a Parallel File System into Hadoop. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Wachs, Matthew. Performance Insulation in Shared Cluster-based Storage. 2008 Parallel Data Laboratory Workshop & Retreat. Nov 3-5, 2008.

Ward, Lee. Efficacy of GPUs in RAID Parity Calculation. Systems Design and Implementation Seminar, Carnegie-Mellon University. September 25, 2008.

Ward, Lee. Some Open Problems in Supercomputing I/O. University of New Mexico's Computer Science Department colloquium. January 29, 2008.

Weil, Sage A. (presenter), Andrew W. Leung, Scott A. Brandt, Carlos Maltzahn, Univ. of California, Santa Cruz. “RADOS: A Fast, Scalable, and Reliable Storage Service for Petabyte-scale Storage Clusters.” Petascale Data Storage Workshop Supercomputing ‘07, Reno, Nevada, Nov. 2007.

FY09

Adams, Ian: "Maximizing Efficiency By Trading Storage for Computation", Workshop on Hot Topics in Cloud Computing (HotCloud '09), San Diego, CA, June 2009.

Amer, Ahmed: "Progressive Parity-Based Hardening of Data Stores", 27th International Performance of Computers and Communication Conference (IPCCC '08), Austin, TX, December 2008.

Bent, John, Garth Gibson, Gary Grider, Ben McClelland, Paul Nowoczynski, James Nunez, Milo Polte, and MeghanWingate. PLFS: A Checkpoint Filesystem for Parallel Applications. In *SC09*, Portland, Oregon, November 2009.

Bent, John. A talk on data scheduling in distributed systems to DADC09 in Munich, Germany.

Bent, John. A talk on PLFS was given to Oak Ridge National Labs

Bent, John. A talk on PLFS was given to Panasas

Bent, John. A talk on PLFS was given to the DOE headquarters

Bent, John. Keynote address at Santa Cruz Systems Research Lab retreat.

Bent, John. One hour lecture on HPC Storage, File Systems, and I/O past, present, and future, and PLFS, at the University of Wisconsin.

Fan, Bin, DiskReduce: RAID for Data-Intensive Scalable Computing. 4th Petascale Data Storage Workshop Supercomputing '09, Portland, OR.

Fu, Bin - Astronomy Application of Map-Reduce: A Distributed Friends-of-Friends Algorithm. Carnegie Mellon University Parallel Data Laboratory Workshop & Retreat, Nov 9-11, 2009, Bedford Springs, PA.

Gibson, Garth, Bin Fan, Swapnil Patil, Milo Polte, Wittawat Tantisiriroj, Lin Xiao. Understanding and Maturing the Data-Intensive Scalable Computing Storage Substrate. Microsoft Research eScience Workshop 2009, Pittsburgh, PA, October 16-17, 2009.

Gibson, Garth. Scalable Storage Research. 2009 Carnegie Mellon University Parallel Data Lab Spring Industry Visit Day, May 6, Pittsburgh, PA.

Greenan, Kevin: "A Spin-Up Saved is Energy Earned: Achieving Power-Efficient, Erasure-Coded Storage", Fourth Workshop on Hot Topics in System Dependability (HotDep '08), San Diego, CA, December 2008.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, Carnegie Mellon.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, DOE.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, PDSI.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, DADC.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, JOWOG.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, UC Santa Cruz.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, Northern New Mexico College.

Grider, Gary. “Petascale Data Storage at Los Alamos National Laboratory”, US/Britain Weapons Complex JOWOG Meeting.

Grider, Gary. HEC FSIO Efforts at the HECURA. BoF at SC09 in Portland, SC09. November, 2009.

Grider, Gary."Update on LANL Data and Information Availability" PDSW ‘09 in Portland, SC09. November, 2009.

Honeyman, Peter. An invited talk on PDSI-related research at the EMC Innovation Conference in Franklin, MA on October 22, 2008.

Kang, Yangwook: "Adding Aggressive Error Correction to a High-Performance Flash File System", 9th ACM/IEEE Conference on Embedded Software (EMSOFT '09), Grenoble, France, October 2009.

Kasick, Michael P., Keith A. Bare, Eugene E. Marinelli III, Jiaqi Tan, Rajeev Gandhi, Priya Narasimhan. System-Call Based Problem Diagnosis for PVFS. Proceedings of the 5th Workshop on Hot Topics in System Dependability (HotDep '09). Lisbon, Portugal. June 2009.

Kasick, Mike - Black-Box Problem Diagnosis in Parallel File Systems. Carnegie Mellon University Parallel Data Laboratory Workshop & Retreat, Nov 9-11, 2009, Bedford Springs, PA.

Krevat, Elie - Seeking Efficient Data-Intensive Computing. Carnegie Mellon University Parallel Data Laboratory Workshop & Retreat, Nov 9-11, 2009, Bedford Springs, PA.

Leung, Andrew: "Scalable Full-Text Search for Petascale File Systems", 2008 Petascale Data Storage Workshop (PDSW 08), Austin, TX, November 2008.

Leung, Andrew: "Spyglass: Fast, Scalable Metadata Search for Large-Scale Storage Systems", 7th USENIX Conference on File and Storage Technologies (FAST '09), San Francisco, CA, February 2009.

Loncaric, Calvin. A talk on Ninjat was given at Argonne National Labs

Long, Darrell: "High-Performance Petascale File Systems", Commissariat a` l'Energie Atomique, Direction des Applications Militaires et Direction des Armes Nucleaires, February 2009.

Long, Darrell: "High-Performance Petascale File Systems", Universite de technologie en sciences des organisations et de la decision de Paris–Dauphine, Paris, France, February 2009.

Long, Darrell: "Using Storage Class Memories to Increase the Reliability of Two-Dimensional RAID Arrays", 17th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2009), London, UK, September 2009.

López, Julio - Data-Intensive eScience at CMU. Carnegie Mellon University Parallel Data Laboratory Workshop & Retreat, Nov 9-11, 2009, Bedford Springs, PA.

López, Julio. Data-Intensive Scalable Computing for Science. NASA Goddard Space Flight Center's Information Science & Technology Colloquium Series. February 2009.

Miller, Ethan: "Accurately Estimating Reliability of Heterogeneous Storage Systems", Vrije University, Amsterdam, Netherlands, June 2009.

Miller, Ethan: "Building Reliable, Efficient Metadata Storage on NVRAM", Hanyang University, Seoul, Korea, October 2008.

Miller, Ethan: "Building Reliable, Efficient Metadata Storage on NVRAM", Samsung, Seoul, Korea, October 2008.

Miller, Ethan: "Challenges in Building Long-Term Archival Storage from Smart Bricks", Baskin School of Engineering Research Review Day, Santa Cruz, CA, October 2009.

Miller, Ethan: "Challenges in Preserving Digital Data for the Long Term", keynote talk, 4th IEEE International Conference on Networking, Architecture, and Storage (NAS 2009), Zhangjiajie, China, July 2009.

Miller, Ethan: "Logan: Automatic Management for Evolvable, Large-Scale, Archival Storage", 2008 Petascale Data Storage Workshop (PDSW 08), Austin, TX, November 2008.

Miller, Ethan: "Managing Devices and Finding Data in Multi-Petabyte Media Archives", The Reel Thing XXII, Association of Moving Image Archivists, Los Angeles, CA, August 2009.

Miller, Ethan: "Pergamum: Evolvable Energy Efficient, Reliable, Disk-Based Archival Storage", University of Twente, Enschede, Netherlands, June 2009.

Miller, Ethan: "Reliable and Efficient Metadata Storage and Indexing Using NVRAM", invited talk at the International Workshop on Large-Scale NVRAM Technology (NVRAMOS08) Workshop, Jeju, Korea, October 2008.

Miller, Ethan: "Spyglass: Fast, Scalable Metadata Search for Large-Scale Storage Systems", National University of Defense Technology, Changsha, China, July 2009.

Patil, Swapnil, Garth A. Gibson, Gregory R. Ganger, Julio Lopez, Milo Polte, Wittawat Tantisiroj, and Lin Xiao. In Search of an API for Scalable File Systems: Under the table or above it? USENIX HotCloud Workshop 2009. June 2009, San Diego CA.

Polte, Milo, Garth Gibson, Carnegie Mellon University, Jay Lofstead, Karsten Schwan, Matthew Wolf, Georgia Institute of Technology, John Bent, Meghan Wingate, Los Alamos Nat Lab, Scott A. Klasky, Qing Liu, Norbert Podhorszki, Oak Ridge Nat Lab, Manish Parashar, Rutgers University. ...And eat it too: High read performance in write-optimized HPC I/O middleware file formats. 4th Petascale Data Storage Workshop (PDSW 09), Portland, OR., November 15, 2009.

Polte, Milo. PLFS: The Parallel Log-Structured File System. 17th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA. Nov 2009.

Sambasivan, Raja - Diagnosing Performance Problems by Comparing System Behaviours. Carnegie Mellon University Parallel Data Laboratory Workshop & Retreat, Nov 9-11, 2009, Bedford Springs, PA.

Sambasivan, Raja. Diagnosing Performance Problems by Comparing System Behaviours. 17th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA. Nov 2009.

Storer, Mark: "Secure Data Deduplication", 4th International Workshop on Storage Security and Survivability (StorageSS 2008), held in conjunction with the 15th ACM Conference on Computer and Communications Security (CCS 2008), Arlington, VA, October 2008.

Tantisiriroj, Wittawat. DiskReduce: Making Room for More Data on DISCs. 17th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA. Nov 2009.

Vasudevan, Vijay, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, Brian Mueller. Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication. SIGCOMM'09, August 17–21, 2009, Barcelona, Spain.

Vasudevan, Vijay, Hiral Shah, Amar Phanishayee, Elie Krevat, David Andersen, Greg Ganger, Garth Gibson. Solving TCP Incast in Cluster Storage Systems. FAST 2009 Work in Progress Report. 7th USENIX Conference on File and Storage Technologies. Feb 24-27, 2009, San Francisco, CA.

Vijayakumar, Karthik, Frank Mueller, Xiaosong Ma, North Carolina State University, Philip C. Roth, Oak Ridge Nat Lab. Scalable I/O Tracing and Analysis. 4th Petascale Data Storage Workshop (PDSW 09), Portland, OR. November 15, 2009.

Wachs, Matthew - Performance Insulation: A Foundation for Storage QoS. Carnegie Mellon University Parallel Data Laboratory Workshop & Retreat, Nov 9-11, 2009, Bedford Springs, PA.

Wachs, Matthew, Gregory R. Ganger.Co-scheduling of Disk Head Time in Cluster-based Storage. 28th International Symposium On Reliable Distributed Systems September 27-30, 2009. Niagara Falls, New York, U.S.A.

Wildani, Avani: "Protecting Against Rare Event Failures in Archival Systems", 17th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2009), London, UK, September 2009.

Xiao, Lin. Reliability Modeling for Large Scale Declustered Storage. 17th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA. Nov 2009.

FY10

Adams, Ian, Ethan L. Miller, Mark W. Storer, "Examining Energy Use in Heterogeneous Archival Storage Systems," Proceedings of the 18th Annual Meeting of the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2010), August 2010

Amer, Ahmed, "Design Issues for a Shingled Write Disk System," Proceedings of the Conference on Mass Storage Systems and Technologies, Incline Village, Nevada: IEEE, May 2010.

Chen, Y., H. Song, R. Thakur, P.C. Roth, and X-H. Sun, "Optimizing Collective I/O with Data Layout Awareness for Parallel Applications," (poster) 2010 High End Computing File Systems and I/O Research and Development Workshop (HEC-FSIO 2010), Arlington, Virginia, August 2010.

Fan, Bin. DiskReduce Analysis. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Gibson, Garth. Shingled Writing for Magnetic Disk Drives. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Grider, Gary. HEC-FSIO Awareness Talks, Colorado School of Mines.

Grider, Gary. HEC-FSIO Awareness Talks, Northern New Mexico College.

Grider, Gary. HEC-FSIO Awareness Talks, University of New Mexico.

Grider, Gary. HEC-FSIO Awareness Talks, University of Wisconsin.

Grider, Gary. HEC-FSIO Awareness Talks, US/Britain Weapons Complex JOWOG Meeting.

Kang, Yangwook, Jingpei Yang, Ethan L. Miller: ”Efficient Object-based Storage Management for Storage Class Memories”, Samsung, San Jose, July 2010.

Long, Darrell D. E. “Self-Adjusting Two-Failure Tolerant Disk Arrays,” Proceedings of the 5th International Workshop on Petascale Data Storage (PDSW10), held in conjunction with SC2010, November 2010.

Long, Darrell: “High-Performance Petascale File Systems,” Université de technologie en sciences des organisations et de la décision de Paris–Dauphine.

Long, Darrell: “Maximizing Efficiency by Trading Storage for Computation,” IBM-Amrita Cloud Symposium, Amrita University, Coimbatore, India, December 2009.

Long, Darrell: High-Performance Petascale File Systems,” Commissariat à l’Energie Atomique, Direction des Applications Militaires et Direction des Ames Nucléaires.

Long, Darrell: Science & Technology Committee, Lawrence Livermore National Laboratory and Los Alamos National Laboratory.

Miller, Ethan: "Semantic Data Placement for Power Management in Archival Storage", Petascale Data Storage Workshop, New Orleans, LA, November 2010.

Miller, Ethan: "Storage Systems Research at UC Santa Cruz", Data Domain, September 2010.

Miller, Ethan: Department of Energy proposal panel.

Pâris, Jehan-François “Using a Shared Storage Class Memory Device to Improve the Reliability of RAID Arrays,” Proceedings of the 5th International Workshop on Petascale Data Storage (PDSW10), held in conjunction with SC2010, November 2010.

Pâris, Jehan-François, “Improving Disk Array Reliability Though Expedited Scrubbing,” Proceedings of the Fifth IEEE International Conference on Networking, Architecture, and Storage (NAS 2010), Macau: IEEE, July 2010, pp. 119–125.

Parker-Wood, Aleatha, Christina Strong, Ethan L. Miller, Darrell D. E. Long, "Security Aware Partitioning for Efficient File System Search", 26th IEEE Symposium on Massive Storage Systems and Technologies: Research Track (MSST 2010), May 2010.

Parker-Wood, Aleatha: "Security Aware Partitioning for Effective File System Search", Oracle, Redwood Shores, CA, August 2010.

Patil, Swapnil. GIGA+: Towards FS Directories with Billion Files. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Polte, Milo. PLFS on Jaguar/Lustre at Scale. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Sambasivan, Raja. Diagnosing Performance Changes by Comparing Request Flows. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Tantisiriroj, Wittawat. RAIDTool: A First Step to RAID 6 in HDFS. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Xiao, Lin. Reliability Modeling for Large Scale Declustered Storage. 18th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, October 25-27, 2010.

Yang, Jingpei: ”Efficient Storage Management for Object-based Flash Memory”, 18th Annual Meeting of the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2010), Miami, FL, August 2010.

FY11

Fan, Bin. DiskReduce Reprise. 19th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, Nov 2012.

Fan, Bin. Provable Load Balancing for Randomly Partitioned Cluster Services. 19th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, Nov 2012.

Gibson, Garth. Shingled Disks and their File Systems: ShingledFS. 19th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, Nov 2012.

Sambasivan, Raja. Diagnosing Performance Changes By Comparing Request Flows. 19th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, Nov 2012.

Tantisiriroj, Wittawat. On the Duality of Data-intensive File System Design: Reconciling HDFS and PVFS. 19th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, Nov 2012.

Xiao, Lin. Scalable Metadata Service in HDFS. 19th Annual Parallel Data Lab Workshop & Retreat, Bedford Springs, PA, Nov 2012.

10 Appendix 3: Supported Personnel

FY07

Carnegie Mellon University (Institute lead) (not all students supported for all quarters)

- Garth Gibson, Associate Professor (Institute PI)
- Anastasia Ailamaki, Assistant Professor
- Gregory Ganger, Professor
- David O'Hallaron, Associate Professor
- Bianca Schroeder, Postdoctoral Researcher, departed to join faculty at University of Toronto
- Alice Zheng, Postdoctoral Researcher, departed to join Microsoft Research
- Milo Polte, PhD student
- Swapnil Patil, PhD student
- Jiri Simsa, PhD student
- Wittawat Tantisiriroj, PhD student
- Mike Mesnier, PhD student
- Elie Krevat, PhD student
- Matthew Wachs, PhD student
- James Hendricks, PhD student
- Raja Sambasivan, PhD student
- Shafeeq Sinnamonhideen, PhD student
- Deepti Chheda, Masters student, graduated
- Shobhit Dayal, Masters student

Lawrence Berkeley National Laboratory / National Energy Research Scientific Computing Center

- William Kramer, PDSI Member PI
- Akbar Maktarani, research staff

Los Alamos National Laboratory

- Gary Grider – PDSI Member PI
- James Nunez – research Staff
- John Bent – research Staff
- Milo Polte – CMU student summer intern at LANL
- Andy Kowinski – Wisconsin student summer intern at LANL

Oak Ridge National Laboratory

- Philip Roth, PDSI Member PI
- Richard Barrett, research staff

Pacific Northwest National Laboratory

- Evan Felix, PDSI Member PI
- Robert Farber, sub-project leader, research staff
- David Brown, research staff

Sandia National Laboratory

- Lee Ward, PDSI Member PI
- Brian Kellogg, research staff
- Ruth Klundt, research staff
- Marlow Weston, research staff
- Matt Curry, Ph.D. Candidate, University of Alabama, Birmingham

University of Michigan

- Peter Honeyman, Scientific Director. PDSI Member PI.
- W.A. (Andy) Adamson, Research Investigator.
- J. Bruce Fields, Asst. Research Scientist.
- Dean Hildebrand, PhD student and Research Assistant. Defended his doctoral thesis in February 2007, but stayed.
- Jiaying Zhang, PhD student and Research Assistant. Defended doctoral thesis in May 2007.
- Alexander Soule, Research Intern, undergraduate Computer Science
- Eva Kramer, Research Intern, undergraduate Electrical Engineering and Computer Science
- Meelap Shah, Research Intern, undergraduate Mathematics and Computer Science
- Nicholas Seltzer, Research Intern, undergraduate Computer Science
- Eaman Jahani, Research Intern, undergraduate Civil Engineering

University of California at Santa Cruz (not all students supported for all quarters)

- Darrell Long, faculty, PDSI Member PI.
- Ethan Miller, faculty
- Scott Brandt, faculty
- Carlos Maltzahn, faculty
- Kevin Greenan, PhD graduate student
- Mark Storer, PhD graduate student
- Jonathan Koren, PhD graduate student
- Sasha Ames, PhD graduate student
- Andrew Leung, PhD graduate student
- Eric Lalonde, masters student, graduated.

FY08

Carnegie Mellon University (Institute lead) (not all students supported for all quarters)

- Garth Gibson, Professor (Institute PI)
- Gregory Ganger, Professor
- David O’Hallaron, Associate Professor
- Julio Lopez, Research Faculty
- Michael Abd-El-Malek, PhD student
- James Hendricks, PhD student
- Mike Kasick, PhD student
- Andrew Klosterman, PhD student
- Elie Krevat, PhD student
- Mike Mesnier, PhD student, defended PhD dissertation 11/27/07, now with Intel, Hillsboro, OR

- Milo Polte, PhD student
- Swapnil Patil, PhD student
- Raja Sambasivan, PhD student
- Shafeeq Sinnamonhideen, PhD student
- Jiri Simsa, PhD student
- Wittawat Tansitiriroj, PhD student
- Matthew Wachs, PhD student
- Shobhit Dayal, Masters student, graduated May 2008, now with Data Domain.

Lawrence Berkeley National Laboratory / National Energy Research Scientific Computing Center

- William Kramer, PDSI Member PI
- Akbar Maktarani, research staff
- Jason Hick, research staff

Los Alamos National Laboratory

- Gary Grider, LANL Staff – PDSI Task PI
- James Nunez, LANL Staff
- John Bent, LANL Staff
- Meghan Quist, LANL Staff
- Andy Nelson, LANL Staff

Oak Ridge National Laboratory

- Philip Roth, PDSI Member PI
- Nirmal Thacker, intern, Georgia Institute of Technology
- Tristin Cuevas, intern, Kent State University

Pacific Northwest National Laboratory

- Evan Felix, PDSI Member PI
- Robert Farber, sub-project leader, research staff
- David Brown, research staff
- Brock Erwin – PNNL Student Researcher
- Arthur Wanner – PNNL Student Researcher

Sandia National Laboratory

- Lee Ward, PDSI Member PI
- Brian Kellogg, research staff
- Ruth Klundt, research staff
- Marlow Weston, research staff
- Matt Curry, Ph.D. Candidate, University of Alabama, Birmingham

University of Michigan

- Dr. Peter Honeyman, Research Professor of Information, U. Mich. PI for PDSI.
- Mr. W.A. (Andy) Adamson, Research Investigator, PDSI-related research and development (left the U. of Michigan in April 2008).
- Dr. J. Bruce Fields, Assistant Research Scientist, PDSI-related research and development

- Mr. James Rees, Systems Analyst, conducts PDSI-related research and development.
- Mr. David Richter, Systems Programmer, conducts PDSI-related research and development.
- Mr. Zongyun Lai, doctoral pre-candidate studying Computer Science.
- Mr. Josef Sipek, is a doctoral pre-candidate studying Computer Science at the University of Michigan.
- Mr. Eamon Jahani, Research Intern, undergraduate studying Computer Science.
- Ms. Eva Kramer, Research Intern, undergraduate studying Electrical Engineering and Computer Science.
- Mr. Meelap Shah, Research Intern, is an undergraduate studying Mathematics and Computer Science.
- Mr. Alexander Soule, Research Intern, undergraduate studying Computer Science.
- Mr. Nicholas Seltzer, Research Intern, undergraduate studying Computer Science.
- Mr. Johann Dahm, Research Intern, undergraduate studying Engineering.

University of California at Santa Cruz (not all students supported for all quarters)

- Darrell Long, faculty, PDSI Member PI.
- Ethan Miller, faculty
- Carlos Maltzahn, faculty
- Kevin Greenan, graduate student
- Mark Storer, graduate student
- Jonathan Koren, graduate student
- Sasha Ames, graduate student
- Andrew Leung, graduate student

FY09

Carnegie Mellon University (Institute lead) (not all students supported for all quarters)

- Garth Gibson, Professor (Institute PI)
- Gregory Ganger, Professor
- David O'Hallaron, Associate Professor
- Julio Lopez, Research Faculty
- Michael Abd-El-Malek, PhD student, defended PhD dissertation Aug. 4/09 (Google)
- James Hendricks, PhD student, defended PhD dissertation July 16/09 (Google)
- Mike Kasick, PhD student
- Andrew Klosterman, PhD student, defended PhD dissertation Aug. 17/09 (Avere Systems)
- Elie Krevat, PhD student
- Milo Polte, PhD student
- Swapnil Patil, PhD student
- Raja Sambasivan, PhD student
- Shafeeq Sinnamonhideen, PhD student
- Jiri Simsa, PhD student
- Wittawat Tansitiriroj, PhD student
- Matthew Wachs, PhD student

Lawrence Berkeley National Laboratory / National Energy Research Scientific Computing Center

- William Kramer, PDSI Member PI, departed LBNL in November 2008
- John Shalf, PDSI Member PI, replaced William Kramer in December 2008

- Akbar Maktarani, research staff
- Jason Hick, research staff

Los Alamos National Laboratory

- Gary Grider, LANL Staff – PDSI Task PI
- James Nunez, LANL Staff
- John Bent, LANL Staff
- Meghan Quist, LANL Staff
- Andy Nelson, LANL Staff
- Calvin Loncaric – LANL summer student

Oak Ridge National Laboratory

- Philip Roth – PDSI Member PI
- Karthik Vijayakumar, intern, North Carolina State University

Pacific Northwest National Laboratory

- Evan Felix – PNNL Staff - PDSI Task PI
- Rob Farber – PNNL Staff
- David Brown – PNNL Staff
- Brock Erwin – PNNL Student Researcher
- Arthur Wanner – PNNL Student Researcher
- Chris Simmons – PNNL Student Researcher
- Aby Kuruvilla – PNNL Masters Student Researcher

Sandia National Laboratory

- Lee Ward, PDSI Member PI
- Ruth Klundt, research staff
- Marlow Weston, research staff

University of Michigan

Faculty (2), research staff (1), graduate students (3), undergraduate students (3).

- Dr. Peter Honeyman, Research Professor of Computer Science and Engineering, is the University of Michigan Principal Investigator for PDSI.
- Dr. J. Bruce Fields, Assistant Research Scientist, conducts PDSI-related research and development at the University of Michigan.
- Mr. David Richter, Systems Programmer, conducted PDSI-related research and development at the University of Michigan until June 2009.
- Mr. Josef Sipek is a doctoral pre-candidate studying Computer Science at the University of Michigan.
- Mr. Eaman Jahani advanced from Research Intern (Fall 2008 through Summer 2009) to doctoral pre-candidate (Fall 2009) studying Computer Science at the University of Michigan following his December 2008 graduation with a B.S., double major in Civil Engineering and in Computer Science and Engineering.
- Mr. Zongyun Lai, a doctoral pre-candidate studying Computer Science at the University of Michigan, was supported by PDSI in Fall 2008.

- Mr. Joseph Maximilian Deliso, Research Intern, is an undergraduate studying Computer Science at Pennsylvania State University.
- Mr. Alexander Soule, Research Intern, graduated in December 2008 with a B.S. in Computer Science and Engineering at the University of Michigan.

University of California at Santa Cruz

Students: (not all students supported for all quarters):

- Ian Adams
- Deepavali Bhagwat
- Kevin Greenan
- Jonathan Koren
- Andrew Leung
- Aleatha Parker-Wood
- Mark Storer
- Christina Strong
- Avani Wildani

Faculty:

- Darrell Long
- Ethan Miller
- Carlos Maltzahn

FY10

Carnegie Mellon University (Institute lead) (not all students supported for all quarters)

- Garth Gibson, Professor (Institute PI)
- Gregory Ganger, Professor
- Julio Lopez, Research Faculty
- Mike Kasick, PhD student
- Elie Krevat, PhD student
- Milo Polte, PhD student
- Swapnil Patil, PhD student
- Raja Sambasivan, PhD student
- Shafeeq Sinnamonhideen, PhD student
- Jiri Simsa, PhD student
- Wittawat Tansitiriroj, PhD student
- Matthew Wachs, PhD student

Lawrence Berkeley National Laboratory / National Energy Research Scientific Computing Center

- John Shalf, PDSI Member PI
- Akbar Maktarani, research staff
- Jason Hick, research staff

Los Alamos National Laboratory

- Gary Grider, LANL Staff – PDSI Task PI
- James Nunez, LANL Staff

- John Bent, LANL Staff
- Meghan Quist, LANL Staff
- Andy Nelson, LANL Staff

Oak Ridge National Laboratory

- Philip Roth – PDSI Member PI
- Yong Chen, ORNL Postdoctoral Research Associate

Pacific Northwest National Laboratory

- Evan Felix, PDSI Member PI
- David Brown
- Brock Erwin

Sandia National Laboratory

- Lee Ward, Principal Investigator
- Marlow Weston, the primary developer, was supported exclusively by this project
- Ruth Klundt, who performed testing and ran the application suites with which we captured data, was partially supported by this project

University of Michigan

Faculty (2), graduate students (2), undergraduate students (4).

- Peter Honeyman, Research Professor of Computer Science and Engineering (University of Michigan Principal Investigator for PDSI).
- J. Bruce Fields, Assistant Research Scientist (PDSI-related research and development at the University of Michigan).
- Josef Sipek, University of Michigan doctoral pre-candidate in Computer Science and Engineering.
- Eaman Jahani, University of Michigan doctoral pre-candidate in Computer Science and Engineering.
- Eric Anderle, Research Intern, University of Michigan undergraduate studying Computer Science and Engineering.
- Michael Groshans, Research Intern, University of Michigan undergraduate studying Computer Science and Engineering.
- Shatarupa Nandi, Research Intern, University of Michigan undergraduate studying Computer Science and Engineering.
- Bryan Smith, Research Intern, University of Michigan undergraduate studying Computer Science and Engineering.

University of California at Santa Cruz (not all students supported for all quarters)

- Darrell Long, faculty
- Ethan Miller, faculty
- Deepavali Bhagwat, student

- Stephanie Jones, student
- Yangwook Kang, student
- Andrew Leung, student
- Alex Nelson, student
- Avani Wildani, student
- Jingpei Yang, student