Final Report
Dynamic Non-Hierarchical File Systems for Exascale Storage
DE-FC02-10ER26017/DE-SC0005417

Darrell D. E. Long (PI)
University of California, Santa Cruz
darrell@cs.ucsc.edu

Ethan L. Miller (Co-PI)
University of California, Santa Cruz
elm@cs.ucsc.edu

February 24, 2015

This is the final report, covering the entire period of funding, including the no-cost time extension (NCTE) through August 2014.

Student Support

We most recently funded three graduate students to work on this project: Stephanie Jones, Christina Strong, and Brian Madden. Other graduate students funded or collaborating on this research include Aleatha Parker-Wood, Yan Li, Michael McThrow, Avani Wildani, Ian Adams, and Preeti Gupta. We are also funded one post-doctoral student, Yasuhiro Ohara. The following students were funded during the Fall 2011 and Winter 2012 quarters to help in specific areas: Alexandra Holloway, Ranjana Rajendran, and Nakul Dhotre.

Stephanie Jones, Christina Strong, and Yan Li have advanced to candidacy and are completing their Ph.D. Brian Madden exited with a M.S. degree and is working at a start-up company in Silicon Valley. Michael McThrow is on leave due to family pressure and is expected to return to finish his Ph.D. Ian Adams, Avani Wildani, and Aleatha Parker-Wood completed their Ph.D. under full or partial funding from this award.

Published Research

Understanding Metadata Needs of HPC Science

Stephanie Jones, Alexandra Holloway, Aleatha Parker-Wood, and Christina Strong traveled to National Laboratories in order to speak with the scientists and learn what their needs and desires were. They visited Los Alamos National Laboratory (LANL), Pacific Northwest National Laboratory (PNNL), and Lawrence Livermore National Laboratory (LLNL). The interviews were focused on the scientists who use the supercomputers at LANL as well as the code developers who write code for use on the supercomputers. These trips directly resulted in two publications in the 2011 Parallel Data Storage Workshop (PDSW '11). One publication is a high level overview touching on what we learned and outlining how our system will help alleviate the problems identified [7]. The other publication is a more detailed look at one of the specific problems that was heard multiple times, that is, how to deal with deleting old data [5].

However, we could only present a small portion of the information we learned while visiting the National Laboratories in the PDSW papers. Since we collected the most information from Los Alamos National Laboratory—the students spent the most time there thanks to the efforts of Meghan Wingate McClelland—we released an technical report including all the anonymized interviews conducted at LANL, compiled from the notes of the students present [17]. The results of these interviews have gained significant interest in the scientific high-performance computing community.

One of the common themes throughout the conversations with scientists was the inability to easily find the data they were looking for. As a result, manual metadata management is common among scientific users, consuming their time while not making use of the computing resources at hand. We proposed a system design that will empower users with more powerful data finding tools, such as unified search spaces, provenance, and ranked file system search [7]. By returning the responsibility of file management, we believe the system should enable scientists to focus on their science without the need for a self-customized file organization scheme for their work.

Another common theme was the "purge threat," as detailed in the second PDSW paper [5]. The three methods currently used to address the purge threat are explained, along with a discussion of how subversion of the purging system is a clear indication of its lack of utility and indicative of its cognitive complexity.

Two file and directory representation models that were proposed to address the purge threat are presented. This paper is an in-depth look at one of the many issues discovered in our discussions, how it is currently handled, and thoughts from the scientists on how it might be handled better.

Data Protection and Security for HPC Storage Systems

During the conversations with scientists, we also learned that security is a largely unanswered problem, and one we had not previously considered. We view this as a critical part of large-scale systems, since data used in high-performance computing (HPC) applications is often sensitive, necessitating protection against both physical compromise of the storage media and "rogue" computation nodes. Preliminary work on this was published in PDSW '11 [15], presenting our approach, Horus, which encrypts petabyte-scale files using a keyed hash tree to generate different keys for each region of the file. This approach supports much needed finer-grained security, since a client can only access a file region for which it has a key, and the tree structure allows keys to be generated for large and small regions as needed.

We then refined our design and implemented a Horus prototype, showing experimentally that the approach works well in providing scalable security for HPC-type workloads [9]. Our Horus prototype employs a client-server model, where the client requests keys from the server (called the key distribution server, or KDS) and uses the keys to encrypt/decrypt each block of a file. Given the root key and the access location for the file, anyone can calculate the leaf key that is actually used to encrypt/decrypt the file. The KDS can be located at any part of the current HPC architecture, placed independently, or clustered as part of a key distribution cluster. Our Horus prototype used a tree of keyed hashes to derive unique encryption keys for individual blocks from a per-file key, resulting in a block-based encryption that only requires one key per file. Lower-level keys can be derived from keys higher in the tree, providing secure access to any size chunk of a file, ranging from one block to the entire file; our experiments showed that this greatly reduces the amount of communication as well as the number of keys that has to be distributed, since the client can derive the per-block keys based on the key they are given. Horus can be integrated into a file system or layered between applications and existing file systems, and poses no added demand on the metadata cluster or the storage devices, and little added demand on the clients, making it highly suitable for protecting data in HPC systems.

At a recent DOE workshop on scientific computing, the need for security and access control such as that provided by Horus was identified as a key requirement.

We have also looked at data protection from a reliability standpoint [8]. To this end we have created Proteus, an open-source simulation program that can predict the risk of data loss in large disk array configurations such as mirrored disks, and all levels of RAID. Proteus characterizes each array by five values, the size of the array, the number of simultaneous disk failures the array will tolerate without data loss, and the respective fractions of failures that will not result in data loss. Our measurements have shown a surprisingly high level of agreement with results obtained via analytical techniques.

Another area we studied was the use of tracing information to reduce power usage or improve reliability by grouping data that is accessed together onto a smaller set of devices [18]. This approach has the potential to reduce power usage by allowing an HPC system to only supply power to some of the devices: those that contain the data that the system is actually using. It can also *improve* reliability by reducing the footprint of a data set. By concentrating data onto a smaller number of devices, it ensures that a storage system failure will impact more files from a single data set, but affect fewer data sets overall.

Lastly, we have continued our scalable reliability research with RESAR: A system for a two-failure tolerant, self-adjusting million disk storage cluster. To address the need for exascale size and efficiency we present RESAR Storage. RESAR employs a reliability scheme that abstracts data management as a graph coloring algorithm, providing a reliability greater than RAID 6 or duplication while only adding a storage overhead of 20%. This technique enables a system to distribute the workload of recovery from disk failure across a system. In our emulations, the work of rebuilding one terabyte of data was evenly distributed across 459 disks and completed in less than four minutes with no service disruptions. In addition our emulations showed that RESAR scaled to one million disks with stable and consistent system behavior.

Our interest in providing reliable storage at the medium to large scale resulted in work on self-repairing disk arrays [11]. As the prices of magnetic storage continue to decrease, the cost of replacing failed disks becomes increasingly dominated by the cost of the service call itself. We propose to eliminate these calls by building disk arrays that contain enough spare disks to operate without any human intervention during their whole lifetime. To evaluate the feasibility of this approach, we have simulated the behavior of two-dimensional disk arrays with n parity disks and $n(n-1)/2$ data disks under realistic failure and repair assumptions. Our conclusion is that having $n(n+1)/2$ spare disks is more than enough to achieve a 99.999 percent probability of not losing data over four years. We observe that the same objectives cannot be reached with RAID level 6 organizations and would require RAID stripes that could tolerate triple disk failures. This work gained a lot of attention, including in the trade press after appearing on `slashdot.org`.

Archival Storage

To better understand the behaviors of large-scale storage system users, we examined access traces from the archive at the National Center for Atmospheric Research's (NCAR) [3, 1]. This work analyzed three years of activities, giving valuable insight into a large-scale scientific archive with over 1600 users, tens of millions of files, and petabytes of data. The examination of system usage at varying levels showed that, while a subset of users were responsible for most of the activity, this activity was widely distributed at the file level. We show that physical grouping of files and directories on media can be used to improve archival system performance, and that with file migrations due to hardware changes the adage of "write-once, read-maybe" is incorrect. Based on these observations, we provided suggestions and guidance for future archive systems as well as suggestions for improved tracing of archival activity.

We are also constructing models to understand the implications of a range of design and external factors on the long-term cost of storing archival data. This work will have significant impact on archival storage management at the National Laboratories, since it will inform decisions that designers of archival storage must make. For example, our first published work explored the impact of increasing up-front cost (capital expense) to reduce operational expenses [4], particularly in the context of using flash instead of disk for long-term storage. We found that, because of the much lower operating cost for flash, it might be cost-effective to use flash rather than disk for archival storage. We are continuing to explore this space, and have a student, Preeti Gupta, whose Ph.D. advancement explored this topic.

Metadata and Indexing

Rounding out our studies of large-scale storage systems, we examined four datasets drawn from HPC research areas. These datasets were studied to determine similarities and differences, and the most effective way to index and store the resulting index data. Findings show that the data is large, heterogeneous and high dimensional even within a discipline, and sparse [14]. These properties introduce challenges for traditional indexing techniques, resulting in a recommendation that new indexing strategies employ column stores for more effective storage.

Finally, we have published work on novel indexing structures for large scale search [10]. Our novel structure, the HCTrie aims to provide intelligent multi-dimensional file search by allowing for searches based on any number of a file's metadata attributes. The HCTrie can utilize the differences across these dimensions to prune the search space, and outperforms MySQL in range searches where not all search dimensions are specified.

In continuing our work on understanding large scale storage systems, we created ExDiff, a tool that uses expectation differencing to validate storage system logs [2]. Our solution can identify development errors such as the omission of a logging point, and runtime errors such as log crashes. ExDiff uses metadata snapshots and activity logs to predict the expected state of the system and compares that with the systems actual state. Mismatches between the expected and actual metadata states can then be used to highlight gaps in log coverage, as well as aid in identifying specific types of missing entries. ExDiff is useful in a number of contexts ranging from research, to validation, to security.

One of the students, Aleatha Parker-Wood, advanced to candidacy in February 2012 and made her advancement proposal available as a technical report [13]. Her advancement proposal addressed the question of how users can quickly find and manage files, without burdening the file system with expensive brute force searches, or requiring the user to become an expert in query languages. It proposes to provide new ranking algorithms which are efficient and effective on large multi-user file systems. In order to reduce the burden of file naming by allowing the system to generate expressive, unique names on the file, the proposal identified a statistical property of data that is likely to select meaningful attributes for file names [12]. Her dissertation, filed in 2014, focuses on looking at ways to improve data management using rich metadata, including provenance. She looked specifically at ranked search, and automating file naming and search result disambiguation for HPC systems. Dr. Parker-Wood recently completed Ph.D., and also conducted postdoctoral research at the Conservatoire National des Arts et Métiers. She is now employed by Symantec Corporation.

Provenance

Efficient provenance storage is an essential step towards the adoption of provenance. But finding an efficient method of compressing provenance data, which can be quite large, has proven elusive. In this paper [19], we analyzed the provenance collected from multiple workloads with a view towards efficient storage. Based on our analysis, we characterized the properties of provenance with respect to long term storage. We then proposed a hybrid scheme that takes advantage of the graph structure of provenance data and the inherent duplication in provenance data. Our evaluation indicates that our hybrid scheme, a combination of web graph compression (adapted for provenance) and dictionary encoding, provides the best tradeoff in terms of compression ratio, compression time and query performance when compared to other compression schemes.

The provenance community has built a number of systems to collect provenance, most of which assume that provenance will be retained indefinitely. However, it is not cost-effective to retain provenance information inefficiently. Since provenance can be viewed as a graph, we note the similarities to web graphs and draw upon techniques from the web compression domain to provide our own novel and improved graph compression solutions for provenance graphs. Our preliminary results show that adapting web compression techniques results in a compression ratio of 2.12:1 to 2.71:1, which we can improve upon to reach ratios of up to 3.31:1 [20].

Information leaks are a constant worry and of particular interest to the National Laboratories as well as private scientific research. Provenance data provides yet another source of valuable information, in particular about workflow and processes. After a leak occurs it is very important for the data owner to not only determine the extent of the leak, but who originally leaked the information. We proposed a technique to extend data provenance to aid in determining potential sources of information leaks [6]. While data provenance is commonly defined as the ancestry of a file, the ancestry recorded depends on the provenance collector. Instead of only recording where a file *came from*, we also track when and where a file *leaves* the system. To track these departures, we suggested the use of *ghost objects* when a file is either written to a mounted external storage device or copied to a client machine via NFS or any other network interface such as SSH or FTP. Our solution tracks emigrant data and demonstrates the minor changes to current provenance-aware storage systems required to enable it.

Continuing Research

Here we detail some of the research that was begun under this award, and has resulted in the advancement to candidacy of several students. These students are expected to complete their Ph.D. over the course of this year, and their accomplishments will be largely due to this award.

Reducing Data Movement

We are swiftly approaching the point at which data movement is more expensive than computing. To this end, we have been exploring ways to reduce data movement. We have two students working on different approaches to the problem; one is looking at initially allocating data in an optimal way, the other is developing intelligent cleaning methods.

Data Allocation

By optimally allocating data initially, we can simultaneously reduce the amount data moves within the system as well as making sure the system is being used to its full potential. Current file systems optimize for a single objective, often to the detriment of other objectives. This can result in an unbalanced system that does not reflect the needs of the system or its users. We approached the data allocation problem as a multi-objective optimization problem, designing a general data allocation framework. The general data allocation framework was intended to provide us with a means of finding an optimal data allocation, tailored to the requirements of individual systems, requiring at most the specification of the relative importance of competing performance goals. For such an approach to be feasible, we need to define meaningful, yet observable, metrics for such a multi-objective optimization problem.

We chose system responsiveness, load balancing, and energy savings as our primary objectives, specifically choosing objectives that don't necessarily compliment each other. We measure system responsiveness as the amount of time it takes a request to complete from the point of view of the user. Load balancing we measure from the system side, where we are attempting to evenly spread the load over all disks in the system. Finally, energy savings is measured by minimizing the total number of disks used, so that some can be spun down or even shut off completely.

We developed observable metrics corresponding to each objective, and evaluated them using data sets provided by Los Alamos National Laboratory; the results can be found in our technical report. However, we had to abandon our metric for load balancing, as the mathematics behind it ended up not being viable in a high performance computing environment. As a result, we have been developing a new model for optimal data and workload allocation, using a combination of ideas from bin packing and queuing theory [16].

Shingled Magnetic Recording Disks

We are also attempting to reduce long term data movement by developing a smarter method of cleaning. We are exploring this in the context of shingled magnetic recording (SMR) disks, as they are unique in that intelligent cleaning is not just a bonus, it is necessary. SMR disks are comprised of disk tracks that overlap to create shingles which are grouped together and divided up into bands. Because shingled disks have these overlapping shingles, blocks can not be overwritten in place. The entire band must be read, and the compacted contents of the band are then written to a new band. This means data that is overwritten leaves holes in the bands.

We propose write heat as a metric to reduce the amount of data moved when performing band compaction. Write heat is measured in the frequency of writes to a LBA. Bands that contain a mix of hot and cold data are more expensive to compact than bands that contain only hot or cold data, and can result in a higher frequency of band compaction. By separating incoming write data based on heat, we can reduce the likelihood of a band containing a mixture of hot and cold data. Data blocks are considered to be "hot" if they have been overwritten at least once and are considered to be "cold" otherwise. Bands that contain only hot data are compacted more often but will result in less data needing to be copied per "hot" band compaction. Bands that contain only cold data are compacted less often, but have more data copied per "cold" band compaction. Classifying data as hot or cold and placing it accordingly helps to reduce both the number of bands read before performing compaction and the amount of data copied during band compaction.

In order to simulate shingled disks, we implemented a simulated log-structured file system (LFS) with a block-based API. We chose to do so because a LFS segment can be treated much like a SMR band, and segment cleaning in LFS is very similar to band compaction for shingled disks. However, LFS can't be directly ported to SMR drives because of the unique characteristics of shingled disks. Therefore, we have developed and tested three different heuristics that define how to select which bands should be compacted; each iteration builds off the previous heuristic. Also, in order to do this separation as early as possible, we implemented a two segment write buffer. However, we found that this absorbs the majority of the write traffic in our current data. Our immediate future work looks at a way to pre-populate the log with data, in order to force band compaction.

Archival Storage Modeling

We are continuing to explore cost tradeoffs in archival storage by constructing a modeling tool that allows us to simulate long-term storage costs based on initial cost assumptions and their change over time. We can model initial and ongoing costs as well as costs associated with migrating data, and can also study the impact of storage density changes and the introduction of new devices, such as DNA-based storage. Our model also includes the ability to ensure that the archive will meet performance goals, such as bandwidth and number of concurrent operations, allowing it to forecast archive storage costs given a wide range of assumptions. We are working with Los Alamos National Laboratory on this project, since they have an interest in preserving scientific data from decades of observations and HPC models. We have also contacted other Department of Energy sites to explore collaboration on this project.

Making Connections

The visits that we made to Los Alamos National Laboratory (LANL), Pacific Northwest National Laboratory (PNNL), and Lawrence Livermore National Laboratory (LLNL) were well received by the interviewees. Since there is a growing trend away from the scientists needing to know how to write their own code, there were many for whom the file system was a "black box" and many were content to leave it that way. Unfortunately, this seemed to result in the scientists managing their own metadata, with techniques to do so ranging from a lab notebook or three ring binder to a PowerPoint presentation or the "Mac Stickies" application. Most of the interviewees, however, were extremely interested in having a better metadata management system, and were not only receptive to what was proposed but also added their own suggestions. The trip to Los Alamos National Laboratory could not have happened without the help of Meghan Wingate McClelland, and the Pacific Northwest National Laboratory visit would not have been possible without the help of John Johnson. We are extremely grateful for the time and effort they put into making these trips successful.

The April 2012 Exascale Meeting in Portland, Oregon enabled us to maintain prior connections, but also to make new ones. Darrell Long attended representing the grant, and Stephanie Jones and Christina Strong attended as student scribes. Having been to the kickoff meeting in 2011, both Stephanie and Christina were glad to have the opportunity to attend again. Both value the perspective they gained at the meetings, since file system design is often an after-thought at this stage. One of the key issues they identified was the need to clearly define terms early on, to eliminate confusion among different fields. In addition, now that everyone is at a more concrete stage than last year, the resulting discussions were incredibly useful to help push forward the file system design.

Unexpended Funds

None.

Research Goals Progress

Below we itemize the research goals we made in the project proposal and how we have addressed each goal. In some places the work has been completed, but in others the word begun under this award has resulted in doctoral research and the students who advanced to candidacy continue the research in order to complete their degree.

- *Integrate dynamic extraction into a file system containing a partition-based metadata server with provenance and content-based metadata.*

  Our initial efforts to incorporate provenance into the metadata server were met with failure, as the provenance collection tool we attempted to use turned out to be experimental code and not functional in real world applications. We discovered that there was no real research into the storage and long term maintenance of data provenance, and have added this as a research focus. As a result, it became necessary to redesign the metadata server. We are exploring ways to keep the cost of dynamic extraction to a minimum while still being able to provide the rich metadata needed.

  It quickly became clear that provenance data was immense, and in some cases may be larger than the actual data itself. As a result, one of our tasks became the management of this data. We developed novel compression techniques.

- *Implement experimental prototypes, and develop initial file system design.*

  After re-working our plan for prototype implementation, we have implemented features to allow Ceph to be utilized as a semantic file system. Adding these features requires that an additional layer be built into Ceph allowing for system calls to be intercepted, and additional code run before, or after the execution of the called function. This new layer will enable future modification with minimal effort, and minimal changes to the Ceph system. In addition to our Ceph modifications, we are also addressing the difficulties of implementing a non-hierarchical file system, as well as the user inertia problem by designing a hybrid file system that seeks to capture the best features of traditional and non-hierarchical systems, while eschewing their failings. Our new system introduces a new encapsulation construct—namespaces—to replace and extend the directory functionality. The resulting system supports both navigation and searching, allowing users to migrate smoothly from traditional hierarchies to non-hierarchical storage while preserving the ability to focus on a subset of the file system.

- *Investigate and develop new approaches for improving metadata management in exascale systems.*

  We are continuing with our efforts to ease the burdens of manual file management by investigating new indexing methods, and policies. We have made good progress with our on-demand indexing policy, and will continue to modify and improve on our initial approach. In addition to the on-demand policy our HCTrie structure shows potential as a new structure for high dimensional indexing. We will continue to investigate how to best integrate this structure into the storage stack, and how it may be most effectively used in HEC environments. Difficulties in obtaining good sample data have slowed our work on search result ranking and automatic naming, however data has been acquired that will allow this work to continue in the coming months.

- *Develop automatic statistics generation and gathering algorithms; refine a statistics generation algorithm for exascale systems and integrate it into the file system.*

  We have worked with the National Laboratories and with several companies in order to obtain dynamic workloads for us analyze. We have developed a set of metrics to measure the popularity, similarity, and coverage of data in a file system. These metrics are based on the statistics gathered and generated from the static file system snapshots we currently have. The metrics are being used in a file allocation algorithm to identify where data should be placed in order to achieve an optimally balanced system.

  We have also done work in statistically identifying how data can be grouped together, and have used that to increase energy savings and reliability in a system. We've also shown that statistical grouping works better than categorical grouping, and with less human intervention required.

- *Evaluate new storage technologies and integration into the file system design.*

  We explored the benefits of using a data management technique that employs multiple logs. When a log fills up, the "cold" data (data used less frequently) is moved to another log. This technique can be used with any underlying storage technology, but we are looking at using it with a flash memory buffer (similar to Gary Grider's "burst buffer") to help amortize the cost of data movement.

- *Evaluate the needs of HEC users and evaluate workflow and interface elements.*

  In order for this to be successful, the scientists need to be able (and willing) to use it. We visited three National Laboratories (LANL, PNNL, and LLNL) this past year, and met with people from ANL and the Norwegian Metacenter for Computational Science (NOTUR) as well. These visits were invaluable during design decisions of the file system, as we were able to learn what the scientists actually need, and what they would like to see. We have stayed in contact with scientists at many of the National Laboratories, as well as the administrators at the National Center for Atmospheric Research (NCAR) and scientists at NOTUR. We intend to continue to get feedback from them as we move forward with implementation. In addition to our previously mentioned efforts, we have also performed our study of scientific data to better understand the needs of HPC scientists. This study will allow us to make informed decisions regarding indexing, search, and interface elements.

Bibliography

[1] Ian Adams, Brian Madden, Joel Frank, Mark W. Storer, and Ethan L. Miller. Usage behavior of a large-scale scientific archive. In *Proceedings of the 2012 International Conference for High Performance Computing, Networking, Storage and Analysis (SC12)*, November 2012.

[2] Ian F. Adams, Mark W. Storer, Avani Wildani, Ethan L. Miller, and Brian A. Madden. Validating storage system instrumentation. In *Proceedings of the 21st International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '13)*, August 2013.

[3] Joel Frank, Ethan L. Miller, Ian Adams, and Daniel Rosenthal. Evolutionary trends in a supercomputing tertiary storage environment. In *Proceedings of the 20th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '12)*, August 2012.

[4] Preeti Gupta, Avani Wildani, Daniel Rosenthal, Ethan L. Miller, Ian Adams, Christina Strong, and Andy Hospodor. An economic perspective of disk vs. flash media in archival storage. In *Proceedings of the 22nd International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '14)*, September 2014.

[5] Alexandra Holloway. The purge threat: Scientists' thoughts on peta-scale usability. In *Proceedings of the 6th Parallel Data Storage Workshop (PDSW '11)*, November 2011.

[6] Stephanie Jones, Christina Strong, Darrell D. E. Long, and Ethan L. Miller. Tracking emigrant data via transient provenance. In *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP '11)*, June 2011.

[7] Stephanie Jones, Christina Strong, Aleatha Parker-Wood, Alexandra Holloway, and Darrell D. E. Long. Easing the Burdens of HPC File Management. In *Proceedings of the 6th Parallel Data Storage Workshop (PDSW '11)*, November 2011.

[8] Hsu-Wan Kao, Jehan-François Pâris, Darrell D. E. Long, and Thomas Schwarz. A flexible simulation tool for estimating data loss risks in storage arrays. In *Proceedings of the 29th IEEE Conference on Mass Storage Systems and Technologies*, May 2013.

[9] Yan Li, Nakul Sanjay Dhotre, Yasuhiro Ohara, Thomas M. Kroeger, Ethan L. Miller, and Darrell D. E. Long. Horus: Fine-grained encryption-based security for large-scale storage. In *Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST)*, February 2013.

[10] Yasuhiro Ohara. HCTrie: A structure for indexing hundreds of dimensions for use in file systems search. In *Proceedings of the 29th IEEE Conference on Mass Storage Systems and Technologies*, May 2013.

[11] Jehan-François Pâris, Ahmed Amer, Darrell D. E. Long, and Thomas Schwarz. Self-repairing disk arrays. In *Proceedings of the Fifth International Workshop on Adaptive Self-tuning Computing Systems (ADAPT)*, January 2015.

[12] Aleatha Parker-Wood, Darrell D. E. Long, Ethan L. Miller, Philippe Rigaux, and Andy Isaacson. A file by any other name: Managing file names with metadata. In *Proceedings of the 7th Annual International Systems and Storage Conference (SYSTOR 2014)*, June 2014.

[13] Aleatha Parker-Wood, Darrell D. E. Long, Ethan L. Miller, Margo Seltzer, and Daniel Tunkelang. Making sense of file systems through provenance and rich metadata. Technical Report UCSC-SSRC-12-01, University of California, Santa Cruz, March 2012.

[14] Aleatha Parker-Wood, Brian A. Madden, Michael McThrow, Darrell D. E. Long, Ian F. Adams, and Avani Wildani. Examining extended and scientific metadata for scalable index designs. In *Proceedings of the 6th Annual International Systems and Storage Conference (SYSTOR 2013)*, June 2013.

[15] Ranjana Rajendran, Ethan L. Miller, and Darrell D. E. Long. Horus: Fine-grained encryption-based security for high performance petascale storage. In *Proceedings of the 6th Parallel Data Storage Workshop (PDSW '11)*, November 2011.

[16] Christina Strong, Ahmed Amer, and Darrell D. E. Long. Building JACK: Developing metrics for use in multi-objective optimal data allocation strategies. Technical Report UCSC-SSRC-14-01, University of California, Santa Cruz, January 2014.

[17] Christina Strong, Stephanie Jones, Aleatha Parker-Wood, Alexandra Holloway, and Darrell D. E. Long. Los Alamos National Laboratory interviews. Technical report, University of California, Santa Cruz, September 2011.

[18] Avani Wildani, Ethan L. Miller, Ian Adams, and Darrell D. E. Long. PERSES: Data layout for low impact failures. In *22th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2014)*, September 2014.

[19] Yulai Xie, Kiran-Kumar Muniswamy-Reddy, Dan Feng, Yan Li, Darrell D. E. Long, Zhipeng Tan, and Lei Chen. A hybrid approach for efficient provenance storage. In *Proceedings of the 2012 International Conference on Information and Knowledge Management Systems (CIKM '12)*, October 2012.

[20] Yulai Xie, Kiran-Kumar Muniswamy-Reddy, Darrell D. E. Long, Ahmed Amer, Dan Feng, and Zhipeng Tan. Compressing provenance graphs. In *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP '11)*, June 2011.