The University of Maine
DigitalCommons@UMaine

University of Maine Office of Research and Sponsored Programs: Grant Reports

Special Collections

6-4-2010

HEC: Collaborative Research: SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing

Yifeng Zhu
Principal Investigator; University of Maine, Orono, yifeng.zhu@maine.edu

Follow this and additional works at: https://digitalcommons.library.umaine.edu/orsp_reports

Part of the Numerical Analysis and Scientific Computing Commons

Recommended Citation
Zhu, Yifeng, "HEC: Collaborative Research: SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing" (2010). *University of Maine Office of Research and Sponsored Programs: Grant Reports*. 277.
https://digitalcommons.library.umaine.edu/orsp_reports/277

This Open-Access Report is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in University of Maine Office of Research and Sponsored Programs: Grant Reports by an authorized administrator of DigitalCommons@UMaine. For more information, please contact um.library.technical.services@maine.edu.

Final Report for Period: 08/2009 - 07/2010 **Submitted on:** 06/04/2010
Principal Investigator: Zhu, Yifeng . **Award ID:** 0621493
Organization: University of Maine
Submitted By:
Zhu, Yifeng - Principal Investigator
Title:
HEC: Collaborative Research: SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing

Project Participants

Senior Personnel

Name: Zhu, Yifeng
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Name: Cai, Zhao
Worked for more than 160 Hours: Yes
Contribution to Project:
2nd year Ph.D. student and Chinese citizen, working on metadata journaling, support started in Jan. 2007

Name: Yue, Jianhui
Worked for more than 160 Hours: Yes
Contribution to Project:
3rd year Ph.D. student and Chinese Citizen,working on distributed metadata management, partially supported by this project.

Name: Shareef, Ali
Worked for more than 160 Hours: No
Contribution to Project:
1st year Ph.D.student and U.S. Citizen, working on energy consumption modeling, partially supported by this project.

Name: Lin, Lin
Worked for more than 160 Hours: No
Contribution to Project:
Metadata prefetching

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

DOE Argonne National Laboratory
Dr Rob Ross and Rajeev Thakur, DOE Argonne National Laboratory. The Collaboration focuses on the parallel I/O benchmarking and testing.

Huazhong University of Science and Techn

Dr. Dan Feng, Wuhan National Laboratory for Optoelectronics and Huazhong University of Science and Technology, China

Fermi National Accelerator Laboratory

We have been in contact with Fermilab to monitor the metadata traffic of their dCache installation, as well as with the local US CMS tier2 site dCache installation.

Sun Microsystems

Dr. Yifeng Zhu has recently received the Academic Excellence Grant (AEG) from the SUN Microsystems, Inc. SUN is donating equipment, valued at $63,960, for eight fully-configured Sun Fire T1000 servers to supplement and strengthen the HECURA research.

Other Collaborators or Contacts

Activities and Findings

Research and Education Activities:

Research Activities:
1. We have initiated monitoring for the eventual creation of traces of dCache usage both here at UNL as well as at FNAL.
2. We are in the process of designing a more scalable metadata server for dCache at UNL and U-Maine.
3. We have been conducting research to design and evaluate a significantly enhanced metadata management scheme, called G-HBA, based on our previous work on HBA, in terms of scalability and adaptivity over existing schemes at UNL and U-Maine.
4. We have been working on the design and implementation of an improved metadata prefetching scheme that is based on a data-mining technique at UNL and U-Maine.
5. We have been researching efficient schemes for metadata fault-tolerance and fault-recovery for significantly improved reliability and availability of metadata services at UNL.
6. We have been developing multi-variable forecasting models to analyze and predict file metadata access patterns at UCF.
7. We have been collecting cluster file system metadata access traces of some scientific benchmarks and applications at UCF.
8. We have been developing decentralized, locality-aware metadata grouping schemes Nexus grouping to facilitate the bulk metadata operations such as prefetching at UCF, UNL and U-Maine.
9. We have been developing scalable and adaptive Bloom filter arrays to enforce load balance and increase scalability for file mapping at UNL, U-Maine and UCF.

Education Activities:
1. The topic related to parallel I/O has been developed and taught in a new graduate-level course ECE 574 Cluster Computing at UMaine.
2. The topics related to reliability and availability of storage systems have been developed and incorporated in four senior/graduate level courses at UNL: CSCE 430/830 Computer Architecture, CSCE 432/832 High-Performance Processor Architecture, CSCE 488 Computer Engineering Professional Development, and CSCE 489 Computer Engineering Senior Design Project.
3. At UNL, Dr. Swanson offered the seminar course titled 'Hadoop and HDFS'.

Findings:

1. RAID-6 significantly outperforms the other RAID levels in disk-failure tolerance due to its ability to tolerate arbitrary two concurrent disk failures in a disk array. The underlying parity array codes have a significant impact on RAID-6's performance. We propose a new XOR-based RAID-6 code, called the Partition Code (P-Code). P-Code is a very simple and flexible vertical code, making it easy to understand and implement. It works on a group of (prime-1) or (prime) disks, and its coding scheme is based on an equal partition of a specified two-integer-tuple set. P-Code has the following properties: (1) it is a Maximum-Distance-Separable (MDS) code, with optimal storage efficiency; (2) it has optimal construction and reconstruction computational complexity; (3) it has optimal update complexity (i.e., the number of parity blocks affected by a single data-block update is minimal). These optimal properties of P-Code are proven mathematically. While X-Code is provably optimal and RDP is proven optimal in computational complexity and storage efficiency, the latter in its current form is not optimal in update complexity. We propose a row parity placement strategy for RDP to help it attain optimal update complexity. P-Code

complements the other two optimal RAID-6 codes, X-code and RDP, to provide a near-full set of optimal RAID-6 configurations of typical disk-array size (e.g., 4-20 disks). That is, for any prime in a typical array size range, P-Code can be deployed for (prime-1) disks optimally, while X-code (or P-Code) and RDP can be respectively deployed for (prime) and (prime+1) disks optimally. Moreover, P-Code's potentially beneficial properties such as the flexible association between the blocks and their labels may find useful applications in distributed environments.

2. User I/O intensity has a significant impact on on-line RAID reconstruction performance by virtue of disk bandwidth contention. Based on this observation, we designed a novel scheme, called WorkOut, to significantly boost RAID reconstruction performance by I/O Workload Outsourcing. WorkOut effectively outsources all write requests and popular read requests originally targeted at the degraded RAID set to a surrogate RAID set during reconstruction. Our lightweight prototype implementation of WorkOut and extensive trace-driven and benchmark-driven experiments demonstrate that, compared with the existing reconstruction approaches, WorkOut significantly speeds up the reconstruction time and average user response time simultaneously. Importantly, WorkOut is orthogonal to and can be easily incorporated into any existing reconstruction algorithms. Furthermore, it is applicable to improving the performance of other background support RAID tasks such as resynchronization and disk scrubbing.

3. One of the challenging issues in performance evaluation of parallel storage systems through synthetic-trace-driven simulation is to accurately characterize the I/O demands of data-intensive scientific applications. We analyze several I/O traces collected from different distributed systems and conclude that correlations in parallel I/O inter-arrival times are inconsistent, either with little correlation or with evident and abundant correlations. Thus conventional Poisson or Markov arrival processes are inappropriate to model I/O arrivals in some applications. Instead, a new and generic model based on the α-stable process is proposed and validated to accurately model parallel I/O burstiness in both workloads with little and strong correlations. This model can be used to generate reliable synthetic I/O sequences in simulation studies. Experimental results show that this model can capture the complex I/O behaviors of real storage systems more accurately and faithfully than conventional models, particularly for the burstiness characteristics in the parallel I/O workloads.

4. A large-scale distributed file system must provide a fast and scalable metadata lookup service. In large-scale storage systems, multiple metadata servers are desirable for improving scalability. To this end we have proposed a novel scheme, called Group-based Hierarchical Bloom Filter Array (G-HBA), judiciously utilizes Bloom filters to efficiently route requests to target metadata servers. Our G-HBA scheme extends the current Bloom filter-based architecture by considering dynamic and self-adaptive characteristics in ultra large-scale file systems. Our scheme logically organizes metadata servers (MDS) into a multi-layered query hierarchy and exploits grouped Bloom filters to efficiently route metadata requests to desired MDSs through the hierarchy. This metadata lookup scheme can be executed at the network or memory speed, without being bounded by the performance of slow disks. Experimental results show that this scheme can significantly improve metadata management scalability and query efficiency in ultra large-scale storage systems.

5. Obtaining representative and concise I/O workloads for the purpose of evaluating the performance of storage systems remains a challenge due to the complex nature of I/O behaviors. Previous studies have shown that disk I/O traffic can be represented as an independent and identically distributed random process in some workloads and a self-similar process in others. Additionally, workloads in the presence of self-similarity can exhibit either Gaussian or non-Gaussian characteristics. We have proposed a new and generic model based on the α-stable process to accurately build a synthetic workload representative of I/O traffic in production storage systems. The novelty of this new model is that it has the capability of characterizing both self-similar Gaussian and non-Gaussian workloads. Experimental results presented show that this model can more accurately capture the complex I/O behaviors of real storage systems than conventional models, particularly the burstiness and heavy-tail distribution under the Gaussian and non-Gaussian workloads.

6. File correlation, which refers to a relationship among related files that can manifest in the form of their common access locality (temporal and/or spatial), has become an increasingly important consideration for performance enhancement in Peta-scale storage systems. Previous studies on file correlations mainly concern with two aspects of files: file access sequence and semantic attribute. Based on mining with regard to these two aspects of file systems, various strategies have been proposed to optimize the overall system performance. Unfortunately, all of these studies consider either file access sequences or semantic attribute information separately and in isolation, thus unable to accurately and effectively mine file correlations, especially in large-scale distributed storage systems. We have developed a novel File Access corRelation Mining and Evaluation Reference model (FARMER) for optimizing Peta-scale file system performance that judiciously considers both file access sequences and semantic attributes simultaneously to evaluate the degree of file correlations by leveraging the Vector Space Model (VSM) technique adopted from the Information Retrieval field. We extract the file correlation knowledge from some typical file system traces using FARMER, and incorporate FARMER into a real large-scale object-based storage system as a case study to dynamically infer file correlations and evaluate the benefits and costs of a FARMER-enabled prefetching algorithm for the metadata servers under real file system workloads. Experimental results show that FARMER can mine and evaluate file correlations more accurately and effectively. More significantly, the FARMER-enabled prefetching algorithm is shown to reduce the metadata operation latency by approximately 24-35% when compared to a state-of-the-art metadata prefetching algorithm and a commonly used replacement policy.

7. We also propose a simple but powerful on-line availability upgrade mechanism, Supplementary Parity Augmentations (SPA), to address the availability issue for parity-based RAID systems. The basic idea of SPA is to store and update the supplementary parity units on one or a few newly augmented spare disks for on-line RAID systems in the operational mode, thus achieving the goals of improving the reconstruction performance while tolerating multiple disk failures and latent sector errors simultaneously. By applying the exclusive OR operations appropriately among supplementary parity, full parity and data units, SPA can reconstruct the data on the failed disks with a fraction of the original overhead that is proportional to the supplementary parity coverage, thus significantly reducing the overhead of data regeneration and decreasing recovery time in parity-based RAID systems. In particular, SPA has two supplementary-parity coverage orientations, SPA Vertical and SPA Diagonal, which cater to user's different availability needs. The former, which calculates the supplementary parity of a fixed subset of the disks, can tolerate more disk failures and sector errors; whereas, the latter shifts the coverage of supplementary parity by one disk for each stripe to balance the workload and thus maximize the performance of reconstruction during recovery. The SPA with a single supplementary-parity disk can be viewed as a variant of but significantly different from the RAID5+0 architecture in that the former can easily and dynamically upgrade a RAID5 system to a RAID5+0-like system without any change to the data layout of the RAID5 system. Our extensive trace-driven simulation study shows that both SPA orientations can significantly improve the reconstruction performance of the RAID5 system while SPA Diagonal significantly improves the reconstruction performance of RAID5+0, at an acceptable performance overhead imposed in the operational mode. Moreover, our reliability analytical modeling and Sequential Monte-Carlo simulation demonstrate that both SPA orientations consistently more than double the MTTDL of the RAID5 system and improve the reliability of the RAID5+0 system noticeably.

8. To improve energy-efficiency, we propose RoLo (Rotated Logging), a novel logging architecture for write-intensive high-end computing that combines rotated logging with decentralized destaging to construct a logical logger with unlimited capacity. By rotating the active logging space among multiple disks, RoLo can effectively decrease energy consumption and increase reliability of replication-based storage systems, while avoiding the potential performance bottleneck and single point of failure of conventional logging architecture with extra dedicated log disks. We develop three flavors of RoLo, RoLo-E/R/P, to emphasize on energy efficiency, reliability, and performance respectively. Extensive trace-driven evaluations show their respective advantages. RoLo-E achieves up to 81.7% energy saving from a typical RAID10 system. RoLo-P provides the best performance among the RoLo schemes in most cases while RoLo-R achieves the highest reliability in terms of combined measure of MTTDL and disk-spin frequency at a performance cost of 3.77%-4.35% and no cost in energy efficiency.

9. We present DEBAR, a scalable and high-performance de-duplication storage system for backup and archiving, to overcome the throughput and scalability limitations of the state-of-the-art data de-duplication schemes, including the Data Domain De-duplication File System (DDFS). DEBAR uses a Two-Phase De-duplication Scheme (TPDS) that exploits memory cache and disk index properties to judiciously turn the notoriously random and small disk I/Os of fingerprint lookups and updates into large sequential disk I/Os, hence achieving a very high de-duplication throughput. The salient feature of this approach is that both the system backup and archiving capacity and the de-duplication performance can be dynamically and cost-effectively scaled up on demand; it hence not only significantly improves the throughput of a single de-duplication server but also is conducive to distributed implementation and thus applicable to large-scale and distributed storage systems.

10. We develop IDEAS, an identity-based security architecture for large-scale and high-performance storage systems, designed to improve security, convenience and total cost of access control by merging identity management with access control in these systems. IDEAS authenticates users at each I/O node by using a single-identity certificate without the service of a centralized security server and enforces access control mechanism by using an Object-Based Access Control (OBAC) model, which is designed to address the complexity and scalability issue of security administration in large-scale storage systems. We also identify the issue of how to identify and authenticate a large number of users with the state-of-the-art cryptographic solutions and suggest the potential alternative technologies to the well-known PKI mechanism. In particular, we present a generic definition and formal description of the OBAC model. The access control rules for OBAC, namely, the PIPS (Proximity, Inheritance, Priority, Sharing) rules, can be used as the basis for establishing a testing and evaluation criteria for securing general large-scale storage systems. Experiments on the IDEAS prototype in implemented base on a object-based storage system show that IDEAS significantly outperforms the conventional capability-based security scheme (CapSec) in terms of latency for key security-related operations, by a speedup factor of 1.81 and 2.22 for the frequent read and write operations respectively and by a factor of 1.65, 1.22, and 0.52 for the infrequent create, delete and chmod operations respectively. Furthermore, in addition to achieving higher security, IDEAS drastically improves scalability by completely removing the performance bottleneck caused by security overhead through avoiding capability requests for both read and write operations, as evidenced by the zero read and write latency of IDEAS on the metadata server while CapSec quickly saturates its metadata server with a moderate number of read or write requests.

11. Prefetching is an effective technique for improving file access performance, which can significantly reduce access latency for I/O systems. In distributed storage systems, prefetching for metadata files is critical for the overall system performance. We have proposed an Affinity-based Metadata Prefetching (AMP) scheme for metadata servers in large-scale distributed storage systems to provide aggressive metadata prefetching. Through mining useful information about metadata accesses from past history, AMP can discover metadata file affinities accurately and intelligently for prefetching. Compared with LRU and some of the latest file prefetching algorithms such as Nexus and C-Miner, our trace-driven simulations show that AMP can improve buffer cache hit rates by up to 12%, 4.5% and 4% respectively, while reduce the

average response time by up to 60%, 12% and 8%, respectively.

12. For parallel I/O workloads, the memory energy efficiency is determined by a complex interaction among four important factors: (1) cache hit rates that may directly translate performance gain into energy saving, (2) cache populating schemes that perform buffer allocation and affect access locality at the chip level, (3) request clustering that aims to temporally align memory transfers from different buses into the same memory chips, and (4) access patterns in workloads that affect the first three factors.We have developed a new energy-aware buffer cache replacement algorithm. Simulation results based on three real-world I/O traces, including TPC-R, web search and Finance applications, show that our algorithms can save up to 65.1% energy with marginal degradation in hit rates.

13. Indirect blocks, part of a file's metadata used for locating this file's data blocks, are typically treated indistinguishably from file's data blocks in buffer cache. We have found that this conventional approach significantly detriments the overall energy efficiency of memory systems. Scattering small but frequently accessed indirected blocks over all memory chips reduce the energy saving opportunities. We propose a new energy-efficient buffer cache management scheme, named MEEP, which separates indirect and data blocks into different memory chips. Our trace-driven simulation results show that our new scheme can save memory energy up to 20.8% for I/O-intensive server workloads.

14. To achieve a reasonably good trade-off among the three important storage design objectives of performance, reliability and energy-efficiency, we have developed an energy efficient disk array architecture, called a Green RAID (or GRAID), which extends the data mirroring redundancy of RAID10 by incorporating a dedicated log disk. The goal of GRAID is to significantly improve energy efficiency or reliability of existing RAID-based systems without noticeably sacrificing their reliability or energy efficiency. Reliability analysis shows that the reliability of GRAID, in terms of MTTDL (Mean Time To Data Loss), is only slightly worse than RAID10 but much better than other existing RAID-based energy optimizing schemes such as EERAID and PARAID by up to 347 times, with an average of 155 times. On the other hand, our prototype implementation of GRAID and indirect comparisons show that GRAID's energy efficiency is significantly better than that of RAID10 by up to 32.1%, with an average of 25.4% while slightly better than or comparable to EERAID and PARAID.

15. Prefetching simply cannot help when the system is overloaded. We studied the scalability of the metadata servers equipped with Nexus prefetching algorithm by simulating large numbers of clients and servers. In the 16-server case, the throughput increases approximately 6% when the number of clients increase from 1000 to 2000, after that it stops growing since the system became saturated. With 64 or 256 servers, the system throughput scales up almost proportionally with the number of clients, indicating near optimal scalability of the system.

16. A lightweight segment structured local file system component named LSFS can be used to boost the local file metadata I/O performance for state-of-the-art parallel file systems. Parallel virtual file system (PVFS2) is chosen as an example for study. LSFS bridges the mapping gap by introducing a novel compact segment I/O technique, which facilitates the large-only raw disk I/O operations with the help of appropriate dynamic grouping algorithms. The current experimental results indicate that an LSFS-enhanced PVFS2 prototype system can significantly outperform a Linux-Ext3-based PVFS2 by up to 130% higher I/O bandwidth.

17. Recent years have seen a fast-growing volume of I/O traffic propagated through the local I/O interconnect bus. This raises up a question for storage servers on how to resolve such a potential bottleneck. We develop a hierarchical Data Cache Architecture called DCA to effectively slash local interconnect traffic and thus boost the storage server performance. A popular iSCSI storage server architecture is chosen as an example. DCA is composed of a read cache in NIC called NIC cache and a read/write unified cache in host memory called Helper cache. The NIC cache services most portions of read requests without fetching data via the PCI bus, while the Helper cache 1) supplies some portions of read requests per partial NIC cache hit, 2) directs cache placement for NIC cache, and 3) absorbs most transient writes locally. We develop a novel State-Locality-Aware cache Placement algorithm called SLAP to improve the NIC cache hit ratio for mixed read and write workloads. To demonstrate the effectiveness of DCA, we develop a DCA prototype system and evaluate it with an open source iSCSI implementation under representative storage server workloads. Experimental results showed that DCA can boost iSCSI storage server throughput by up to 121 percent and reduce the PCI traffic by up to 74 percent compared with an iSCSI target without DCA.

18. Most of existing search algorithms for unstructured peer-to-peer (P2P) systems share one common approach: the requesting node sends out a keyword search query and the query message is repeatedly routed and forwarded to other peers in the overlay network. Due to multiple hops involved in query forwarding, the search may result in a long delay before it is answered. Furthermore, some incapable nodes may be overloaded when the query traffic becomes intensive or bursty. In this work, we develop a novel content-pushing, Advertisement-based Search Algorithm for unstructured Peer-to-peer systems (ASAP). An advertisement (ad) is a synopsis of contents a peer tends to share, and appropriately distributed and selectively cached by other peers in the system. In ASAP, nodes proactively advertise their contents by delivering ads, and selectively storing interesting ads received from other peers. Upon a request, a node can locate the destination nodes by looking up its local ads repository, and thus obtain a one-hop search latency with modest search cost. Comprehensive experimental results show that, compared with traditional query-based search algorithms, ASAP achieves much better search efficiency, and maintains system load at a low level with small variations. In addition, ASAP works well under node churn.

Training and Development:

The graduate students, Jianhui Yue, Zao Cai, and Ali Shareef, have made significant progress in their research and started publishing research papers in premier venues.

Outreach Activities:

The research results are being incorporated into two NSF-funded education projects. Both projects are leaded by Dr. Yifeng Zhu, the PI of this project. These two projects are (1) NSF-REU SuperMe Project that provides ten-summer-week supercomputing research experiences for 30 undergraduates during the project lifetime, and (2) NSF-ITEST IDEAS Project that allows 60 middle-school teachers and 180 middle-school students during the project lifetime to perform numeric modeling on the supercomputer housed at the University of Maine as well as their laptops in the middle-school classrooms.

Journal Publications

Y. Zhu, H. Jiang, J. Wang, and F. Xian, "HBA: A Distributed Metadata Management System for Large Cluster-based Storage", IEEE Transaction on Distributed and Parallel Computing, p. 750-763, vol. 19, no., (2008). Published,

Y. Zhu and H. Jiang, "RACE: A Robust Adaptive Caching Strategy for Buffer Cache", IEEE Transaction on Computers, p. 25-40, vol. 57, no., (2008). Published,
http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/trans/tc/&toc=comp/trans/tc/2008/01/ttc200801toc.xml&DOI=10.1109/TC.200

Peng Gu, Jun Wang, Yifeng Zhu, Hong Jiang, "Improving Metadata Server Performance in Petabyte-Scale File Systems", Journal of Parallel and Distributed Computing, p. , vol. , (2008). Submitted,

Yu Hua, Dan Feng, Hong Jiang, and Lei Tian, "RBF: A New Storage Structure for Space-Efficient Queries for Multidimensional Metadata in OSS", the 5th USENIX Conference on File and Storage Technologies (FAST '07) Work-in-Progress (WiP) Report, p. 1, vol. , (2007). Published,

Lei Tian, Dan Feng, Hong Jiang, Ke Zhou, Lingfang Zeng, Jianxi Chen, Zhikun Wang, and Zhenlei Song, "PRO: A Popularity-based Multi-threaded Reconstruction Optimization for RAID-Structured Storage Systems", Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), p. 277, vol. , (2007). Published,

Y. Zhu and H. Jiang, "On the Analysis and Impact of False Rates of Bloom Filters in Distributed Systems", Proceedings of the 35th International Conference on Parallel Processing (ICPP), p. 255, vol. , (2006). Published,

P. Gu, Y. Zhu, H. Jiang, and J. Wang, "Nexus: A Novel Weighted-Graph-Based Prefetching Algorithm for Metadata Servers in Petabyte-Scale Storage Systems", Proceedings of International Symposium on Cluster Computing and the Grid (CCGrid, 2006), p. 409, vol. , (2006). Published,

J. Yue, Y. Zhu and Z. Cai, "An Energy-Oriented Evaluation of Buffer Cache Algorithms Under Parallel I/O Workloads", IEEE Transaction on Distributed and Parallel Computing, p. 1565, vol. 19, (2008). Published,

Peng Gu, Jun Wang, Yifeng Zhu, Hong Jiang, "Improving Metadata Server Performance in Petabyte-Scale File Systems", Journal of Parallel and Distributed Computing, p. , vol. , (2008). Published,

Peng Xia, Dan Feng, Hong Jiang, Lei Tian, and Fang Wang, "FARMER: A Novel Approach to File Access Correlation Mining And Evaluating Reference Model for Optimizing Peta-Scale File Systems Performance", Proceedings of the 17th ACM/IEEE International Symposium on High Performance Distributed Computing (HPDC 2008) (Acceptance rate: 17%), p. , vol. , (2008). Published,

L. Lin, M. Li, H. Jang, and Y. Zhu, "AMP: An Affinity-based Metadata Prefetching Scheme in Large-Scale Distributed Storage Systems", Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'08), p. , vol. , (2008). Published,

J. Yue, Y. Zhu and Z. Cai, "Evaluating Memory Energy Efficiency in Parallel I/O Workloads", Proceedings of IEEE International Conference on Cluster Computing, p. , vol. , (2007). Published, (Best Paper Award)

Peng Gu, Jun Wang, Hailong Cai, "ASAP: An Advertisement-based Search Scheme for Unstructured Peer-to-Peer Systems", International Conference on Parallel Processing (ICPP 2007), p. , vol. , (2007). Published,

Zhongying Niu, Ke Zhou, Dan Feng, Hong Jiang, Frank Wang, Hua Chai, Wei Xiao,and Chunhua Li, "Implementing and Evaluating Security Controls for an Object-Based Storage System", Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST?07) (acceptance rate: 18.7%), p. , vol. , (2007). Published,

Lei Tian, Hong Jiang, Dan Feng, Qin Xin, and Xing Shu, "Implementation and Evaluation of a Popularity-Based Reconstruction Optimization Algorithm in Availability-Oriented Disk Arrays", Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST?07) (acceptance rate: 37%), p. 233, vol. , (2007). Published,

Jun Wang, Peng Gu, Hailong Cai, "An Advertisement-based Peer-to-Peer Search Scheme", Journal of Parallel and Distributed Computing, p. , vol. , (2009). Accepted,

Jun Wang, Xiaoyu Yao, Christopher Mitchell, and Peng Gu, "A hierarchical data cache architecture for iSCSI storage server", IEEE Transactions on Computers, p. 1, vol. 58, (2009). Published,

Yu Hua, Yifeng Zhu, Hong Jiang, Dan Feng and Lei Tian, "Supporting Scalable and Adaptive Metadata Management in Ultra Large-scale File Systems", IEEE Transactions on Parallel and Distributed Systems, p. , vol. , (2009). Submitted,

Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao, "WorkOut: I/O Workload Outsourcing for Boosting RAID Reconstruction Performance", IEEE Transactions on Parallel and Distributed Systems, p. , vol. , (2009). Submitted,

Lei Tian, Hong Jiang, Dan Feng, Qiang Cao, Changsheng Xie, and Qin Xin, "SPA: On-Line Availability Upgrades for Parity-based RAIDs through Supplementary Parity Augmentations", IEEE Transactions on Parallel and Distributed Systems, p. , vol. , (2009). Submitted,

Hailong Cai, Ping Ge, Jun Wang, "Applications of Bloom Filters in Peer-to-peer Systems: Issues and Questions", Proceedings of International Conference on Networking, Architecture, and Storage, p. , vol. , (2008). Published,

Saba Sehrish, and Jun Wang, "Smart Read/Write for MPI-IO", In the 14th International Workshop on High-Level Parallel Programming Models and Supportive Environments, in conjunction with the 23rd IEEE International Parallel and Distributed Processing Symposium, p. , vol. , (2009). Published,

Chao Jin, Hong Jiang, Dan Feng, Lei Tian, "P-Code: A New RAID-6 Code with Optimal Properties", in the Proceedings of the 23rd ACM International Conference on Supercomputing (ICS 09), p. , vol. , (2009). Accepted,

Hui Tian, Ke Zhou, Hong Jiang, Yongfeng Huang, Jin Liu, and Dan Feng, "An Adaptive Steganography Scheme for Voice over IP", in the Proceedings of the 2009 IEEE International Symposium on Circuits and Systems (ISCAS 09), p. , vol. , (2009). Accepted,

Hui Tian, Ke Zhou, Hong Jiang, Yongfeng Huang, Jin Liu, and Dan Feng, "An M-Sequence Based Steganography Model for Voice over IP", in the Proceedings of the 2009 IEEE International Conference on Communications (ICC 09), p. , vol. , (2009). Accepted,

Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao, "WorkOut: I/O Workload Outsourcing for Boosting the RAID Reconstruction Performance", Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST '09), p. , vol. , (2009). Published,

Peng Gu, Jun Wang, Robert Ross, "Bridging the Gap Between Parallel File Systems and Local File Systems: A Case Study with PVFS", Proceedings of ICPP, p. , vol. , (2008). Published,

Y. Hua, H. Jiang, Y. Zhu, D. Feng, L. Tian, "SmartStore: A New Metadata Organization Paradigm with Semantic-Awareness", 7th USENIX Conference on File and Storage Technologies, Work-In-Progress, p. , vol. , (2009). Published,

Y. Zhu and H. Jiang, "Efficient Update Control of Bloom Filter Replicas in Large-Scale Distributed Systems", Book chapter, Handbook of research on scalable computing technologies, p. , vol. , (2009). Accepted,

D. Feng, Q. Zou, H. Jiang, and Y. Zhu, "A Novel Model for Synthesizing Parallel I/O Workloads in Scientific Applications", Proceedings of IEEE International Conference on Cluster Computing, p. 252, vol. , (2008). Published,

Q. Zou, D. Feng, Y. Zhu, H. Jiang, X. Ge, and Z. Zhou, "A Novel and Generic Model for Synthesizing Disk I/O Traffic Based on The Alpha-stable Process", Proceedings of 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 08), p. , vol. , (2008). Published,

Bo Mao, Dan Feng, Hong Jiang, Suzhen Wu, Jianxi Chen, Lingfang Zeng, "GRAID: A Green RAID Storage Architecture with Improved Energy Efficiency and Reliability", Proceedings of the 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 08), p. , vol. , (2008). Published,

J. Yue, Y. Zhu, Z. Cai, "Energy Efficient Buffer Cache Replacement", Proceedings of 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, p. , vol. , (2008). Published,

J. Yue, Y. Zhu, and Z. Cai, "Impacts of Indirect Blocks on Buffer Cache Energy Efficiency", Proceedings of the 37th International Conference on Parallel Processing (ICPP 08), p. , vol. , (2008). Published,

Y. Hua, Y. Zhu, H. Jiang, D. Feng, and L. Tian, "Scalable and Adaptive Metadata Management in Ultra Large-Scale File Systems", Proceedings of the 28th International Conference on Distributed Computing Systems (ICDCS'08), p. , vol. , (2008). Published,

Xiao Qin and Hong Jiang, "Dynamic Load Balancing for I/O-Intensive Applications on Clusters", ACM Transactions on Storage, p. 9, vol. 5, (2009). Published,

Jun Wang, Peng Gu, Hailong Cai, "An Advertisement-based Peer-to-Peer Search Scheme", Journal of Parallel and Distributed Computing, p. 638, vol. 69, (2009). Published,

Peng Gu, Jun Wang, Yifeng Zhu, Hong Jiang, Pengju Shang, "A Novel Weighted-Graph-Based Grouping Algorithm for Metadata Prefetching", IEEE Transactions on Computers, p. 1, vol. 59, (2010). Published,

Xiao Qin, Hong Jiang, Adam Manzanares, Xiaojun Ruan, Shu Yin, "Communication-Aware Load Balancing for Parallel Applications on Clusters", IEEE Transactions on Computers, p. , vol. 59, (2010). Published,

Yu Hua, Yifeng Zhu, Hong Jiang, Dan Feng, and Lei Tian, "Supporting Scalable and Adaptive Metadata Management in Ultra Large-scale File Systems", IEEE Transactions on Parallel and Distributed Systems, p. , vol. , (2010). Accepted,

X. Qin, H. Jiang, A. Manzanares, X.-J Ruan, and S. Yin, "Dynamic Load Balancing Support for I/O-Intensive Jobs in Homogeneous and Heterogeneous Clusters of Workstations", IEEE Transactions on Computers, p. , vol. , (2010). Accepted,

Saba Sehrish and Jun Wang, "Reduced Function Set Architecture for MPI-IO", Journal of Supercomputing, p. , vol. , (2010). Accepted,

Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, and Lei Tian, "Semantic-Aware Metadata Organization Paradigm in Next-Generation File Systems", IEEE Transactions on Parallel and Distributed Systems, p. , vol. , (2010). Submitted,

Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao, "mproving Availability of RAID-Structured Storage Systems by Workload Outsourcing", IEEE Transactions on Computers, p. , vol. , (2010). Submitted,

Lei Tian, Hong Jiang, Dan Feng, Qiang Cao, Changsheng Xie, and Qin Xin, "SPA: On-Line Availability Upgrades for Parity-based RAIDs through Supplementary Parity Augmentations", ACM Transactions on Storage, p. , vol. , (2010). Submitted,

Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, "Rapport: Semantic-sensitive Namespace Management in Large-scale File Systems", The 23rd Annual Supercomputing Conference -- SC'10, p. , vol. , (2010). Submitted,

Zhichao Yan, Hong Jiang, Dan Feng, Lei Tian, Yujuan Tan, and Jingning Liu, "SUV-TM: A Novel Single-Update Version-Management

Scheme for Hardware Transactional Memory Systems", The 16th International Conference on High Performance Computing, p. , vol. , (2010). Submitted,

Zhongying Niu, Hong Jiang, Ke Zhou, Dan Feng, "DSFS: Decentralized Security for Large Parallel File Systems", The 11th ACM/IEEE International Conference on Grid Computing (Grid 2010), p. , vol. , (2010). Submitted,

Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, and Guohui Zhou, "SAFE: A Semantic-Aware Source De-duplication Framework for Efficient cloud backup and restore", Software: Practice and Experience, p. , vol. , (2010). Submitted,

Jian Hu, Hong Jiang, Lei Tian, Lei Xu, "PUD-LRU: An Erase-Efficient Write Buffer Management Algorithm for Flash Memory SSD", the Proceedings of The 18th Annual Meeting of the IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS10), p. , vol. , (2010). Accepted,

Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, and Zhichao Yan, "SAM: A Semantic-Aware Multi-Tiered Source De-duplication Framework for Cloud Backup", the Proceedings of The 39th International Conference on Parallel Processing, p. , vol. , (2010). Accepted,

Yang Hu, Hong Jiang, Dan Feng, Lei Tian, Sshuping Zhang, Jingning Liu, Wei Tong, "Achieving Page-Mapping FTL Performance at Block-Mapping FTL Cost by Hiding Address Translation", Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010), p. , vol. , (2010). Published,

Jiansheng Wei, Hong Jiang, Ke Zhou, Dan Feng, "AD2: A Scalable High-Throughput Exact Deduplication Approach for Network Backup Services", Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010), p. , vol. , (2010). Published,

Yinliang Yue, Hong Jiang, Lei Tian, Fang Wang, Dan Fang, and Quan Zhang, "RoLo: A Rotated Logging Storage Architecture for Enterprise Data Centers", Proceedings of The 30th International Conference on Distributed Computing Systems (ICDCS 2010), p. , vol. , (2010). Published,

Dongyuan Zhan, Hong Jiang, and Sharad Seth, "Exploiting Set-Level Non-Uniformity of Capacity Demand to Enhance CMP Cooperative Caching", Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), p. , vol. , (2010). Published,

Tianming Yang, Hong Jiang, Dan Feng, Zhongying Niu,Ke Zhou, andYaping Wan, "DEBAR: A Scalable High-Performance De-duplication Storage System for Backup and Archiving", Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), p. , vol. , (2010). Published,

Bo Mao, Hong Jiang, Dan Feng, Suzhen Wu, "HPDA: A Hybrid Parity-based Disk Array for Enhanced Performance and Reliability", Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), p. , vol. , (2010). Published,

J. Yue, Y. Zhu, Z. Cai, L. Lin, "Energy and Thermal Aware Buffer Cache Replacement Algorithm", Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), p. , vol. , (2010). Published,

Q. Zou, Y. Zhu and D. Feng, "A study of Self-similarity in Parallel I/O Workloads", Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), p. , vol. , (2010). Published,

Suzhen Wu, Dan Feng, Hong Jiang, Bo Mao, Lingfang Zeng, and Jianxi Chen, "JOR: A Journal-guided Reconstruction Optimization for RAID-Structured Storage Systems", Proceedings of the Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09), p. , vol. , (2009). Published,

Hui Tian, Ke Zhou, Hong Jiang, Dan Feng, "Digital Logic Based Encoding Strategies for Voice-over-IP Steganography", Proceedings of the ACM Multimedia 2009 conference, p. , vol. , (2009). Published,

Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, and Lei Tian, "SmartStore: A New Metadata Organization Paradigm with Metadata

Semantic-Awareness for Next-Generation File Systems", Proceedings of The 22nd International Conference on High Performance Computing, Networking, Storage and Analysis (The 22nd Annual Supercomputing Conference -- SC'09), p. , vol. , (2009). Published,

Saba Sehrish, Jun Wang, and Rajeev Thakur, "Self-detecting Locks to Support MPI-IO Atomicity", EuroPVM/MPI, p. , vol. , (2009). Published,

Saba Sehrish, and Jun Wang, "Smart Read/Write for MPI-IO", the 14th International Workshop on High-Level Parallel Programming Models and Supportive Environments, in conjunction with the 23rd IEEE International Parallel and Distributed Processing Symposium, p. , vol. , (2009). Published,

Chao Jin, Hong Jiang, Dan Feng, Lei Tian, "P-Code: A New RAID-6 Code with Optimal Properties", the Proceedings of the 23rd ACM International Conference on Supercomputing, p. , vol. , (2009). Published,

Books or Other One-time Publications

Web/Internet Site

URL(s):
http://www.eece.maine.edu/research/sam2/
Description:

Other Specific Products

Contributions

Contributions within Discipline:
Scientific computing brings revolution to basic scientific and engineering research. Through simulation rather than physical experiments, new knowledge can be discovered and new technology can be developed more quickly, efficiently and cost-effectively in many areas, especially where experiments are expensive, hazardous, or even impossible to perform or observe. However, data I/O tends to be the performance bottleneck for many scientific simulations. This research project aims to eliminate one of the major bottlenecks in data I/O operations: metadata bottlenecks. The following highlights our research outcomes.

1. Hierarchical directory do not meet scalability and functionality requirements for the next-generation extra-large storage systems. We design a semantic-aware organization, called SmartStore, which exploits metadata semantics of files to limits metadata complex queries to a single or a minimal number of semantically related groups. Experimental results show that Smartstore improves system scalability and reduces query latency over basic database approaches by one thousand times (SC?2009).

2. In extra-large storage systems, users often need to manually navigating hierarchies of billions of files or directories to locate target data. We develop a new metadata management scheme, called Rapport, use semantic correlations, instead of file names, to represent a file. It makes use of semantic correlation to create a semantic-sensitive namespace, which consists of the most closely correlated files identified by using a simple and fast LSH-based computation. Experimental results show that Rapport is very scalable and efficient (submitted to SC?2010).

3. We exploit the semantic information to improve the storage systems. Particularly, we design two de-duplication frameworks, called SAFE and MAD2, for high-throughput backup and restore. SAFE combines the global file-level de-duplication and local chunk-level de-duplication to achieve an optimal tradeoff between the de-duplication efficiency and de-duplication overhead to achieve a short backup time. MAD2 eliminates duplicate data both at the file level and at the chunk level by employing four techniques to accelerate the de-duplication process and evenly distribute data. Experiments shows that MAD2 can achieve at least 100 MB/s for each storage component, and SAFE can shorten the backup times of the existing solutions by an average of 38.7%, and reduce the restore times by a ratio of up to 9.7 : 1. (MSST?2010).

4. A significant extension of our HBA scheme and its prototype implementation that were completed during the first year of the funding, G-HBA, for name space management of ultra large-scale file systems has been developed and prototyped as well (ICDCS?08). The G-HBA

scheme is significantly more scalable, adaptive, and space-efficient than HBA.

5. A novel file correlation mining and evaluation model, called FARMER (HPDC?08), has been developed to accurately and efficiently explore and exploit file correlations to optimize storage system performances. To the best of our knowledge, this is the first such scheme that exploits the coordination and correlation between dynamic semantics of file popularity (access patterns) with static semantics of files to accurately mine file correlations.

6. A test framework has been set up to gather information about metadata access behaviors in dCache (implemented at UNL?s tier-2 CMS site), and substantial redesign and streamlining have been done to the current dCache metadata management to enhance the scalability and reliability of the dCache metadata management. Preliminary results have been promising.

7. We propose a new energy-efficient buffer cache management scheme, named MEEP, which separates indirect and data blocks into different memory chips. Our trace-driven simulation results show that our new scheme can save memory energy up to 20.8% for I/O-intensive server workloads.

8. An extended Parallel Virtual File System prototype system with enhanced local I/O component has been developed and tested on several medium scaled cluster systems. Comprehensive experimental results indicate the system throughput can be boosted by up to 132% using several parallel I/O benchmarks.

9. A substantial fault-recovery mechanism, WorkOut is proposed and developed to boost the reconstruction performance for RAID systems (FAST?09). More importantly, WorkOut is orthogonal to and can be easily incorporated into any existing reconstruction algorithms. Furthermore, it is also applicable to improving the performance of other background support RAID tasks such as re-synchronization and disk scrubbing.

10. A novel RAID-6 coding scheme with optimal properties of optimal storage efficiency, construction and reconstruction computational complexity, and update complexity, P-Code is proposed to resist from double disk failures efficiently (ICS ?09). P-Code is a very simple and flexible vertical code, making it easy to understand, implement, and deploy.

11. R-tree with Bloom Filters (RBF) is being developed. A New Storage Structure for Space-Efficient Queries for Multidimensional Metadata in OSS (WiP of FAST07). We propose a new space-efficient storage structure, called the R-tree with Bloom Filters (RBF), to store multidimensional metadata and achieve point and range query with low operational complexity. The basic idea of our RBF is to expand the classical R-tree to incorporate space-efficient Bloom filters in R-tree nodes, maintaining multidimensional range information and achieving space efficiency.

12. We are improving metadata reliability through popularity and locality based reconstruction schemes (work in progress; preliminary work on file data (read most) was recently presented at FAST?07);

13. Appropriate interdisciplinary use of statistical models is being identified to analyze the metadata access patterns of both local file systems and cluster-based file systems.

Contributions to Other Disciplines:

1. The research project has started to integrate into the US CMS research facility at the University of Nebraska (UNL). US CMS is a collaboration US scientists participating in the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland. UNL's US CMS Tier-2 site is a child site of the Tier-1 site at Fermi Nation Laboratory (FNAL). CMS sites employ dCache, a distributed storage data caching system, to support data access and transfer. We have started to prototype SAM2 toolkit, in particular the prefetching algorithms, into dCache to improve the I/O performance. In addition to CMS Tier-2 facility, UNL's Research Computing Facility (RCF), the primary computation resource at UNL, will benefit from our SAM2 toolkit. RCF includes (1) a distributed-memory supercomputer named Prairiefire that has 256 AMD Opteron processors capable of 88.5Gflops and (2) a shared-memory supercomputer from SGI named Homestead that contains 32 500MHz MIPS processor.

2. At the University of Maine, Dr. Yifeng Zhu is working with researchers in Marine Sciences to alleviate the I/O bottleneck for their simulations. The ocean model developed at UMaine is I/O intensive. We are working to utilize parallel I/O to speed up their applications.

Contributions to Human Resource Development:

At the University of Central Florida, Peng Gu defended his Ph.D. dissertation in June 2008.

At the University of Maine, the research findings and concepts are being incorporated into two innovative NSF-funded education programs, leaded by Dr. Yifeng Zhu, to provide college undergraduates as well as middle-school teachers and their students' firsthand experiences in scientific computing. (1) The Supercomputing Undergraduate Program in Maine (SuperMe), funded by a $300,000 grant from NSF, is an opportunity for 10 UMaine undergraduate students to spend the summer conducting the kind of sophisticated, meaningful scientific research that is usually reserved for more advanced students. (2) With a separate $1.2 million NSF grant, another three-year program aims to integrate supercomputer modeling into the Maine middle-school science curriculum. Called Inquiry-based Dynamic Earth Applications of Supercomputing (IDEAS), the program will allow 20 middle-school teachers and 60 of their students each year to explore the myriad intricacies of UMaine's climate computer model by accessing the supercomputer with their state-issued laptops.

Contributions to Resources for Research and Education:

At the University of Central Florida, Peng Gu defended his Ph.D. dissertation in June 2008.

At the University of Maine, the research findings and concepts are being incorporated into two innovative NSF-funded education programs, leaded by Dr. Yifeng Zhu, to provide college undergraduates as well as middle-school teachers and their students? firsthand experiences in scientific computing. (1) The Supercomputing Undergraduate Program in Maine (SuperMe), funded by a $300,000 grant from NSF, is an opportunity for 10 UMaine undergraduate students to spend the summer conducting the kind of sophisticated, meaningful scientific research that is usually reserved for more advanced students. (2) With a separate $1.2 million NSF grant, another three-year program aims to integrate supercomputer modeling into the Maine middle-school science curriculum. Called Inquiry-based Dynamic Earth Applications of Supercomputing (IDEAS), the program will allow 20 middle-school teachers and 60 of their students each year to explore the myriad intricacies of UMaine's climate computer model by accessing the supercomputer with their state-issued laptops.

Contributions Beyond Science and Engineering:

Conference Proceedings

Categories for which nothing is reported:

Any Book

Any Product

Contributions: To Any Beyond Science and Engineering

Any Conference

Joint Final Report for Award CCF 0621526 (UNL) / 0621493 (UMaine)
08/2006 – 06/2010

1. **Project Title:**
   SAM^2 Toolkit: Scalable and Adaptive Metadata Management for High-End Computing

2. **Project Participants**

1.1 Participant Individuals (*):

(a) University of Nebraska – Lincoln (UNL), Award #: 0621526

Graduate Research Assistants:

1. Brian Bockleman, 4th year Ph.D. student at UNL and US citizen, working on dCache, partial support started in Jan. 2007. Graduated in Summer 2008.
2. Ranjini Srinivas, MS student at UNL and Indian citizen, working on dCache, support started in Jan. 2008. Graduation expected in Fall 2010.
3. Dongyuan Zhan, 2nd year Ph.D. student at UNL and Chinese citizen, working on metadata security, partial support started in August 2007.
4. Jian Hu, 2nd year Ph.D. student at UNL and Chinese citizen, working on reliability and scalability of metadata management and SSD, partial support started in June 2008.
5. Lei Xu, 1st year Ph.D. student at UNL and Chinese citizen, working on distributed file systems, with focus on the Hadoop file system and its metadata management, partial support started in January 2009.

Scholar

1. Dr. Lei Tian, a visiting scholar from China at UNL, who works on metadata and file data reliability and power-efficiency, support started in October 2007 and ended in March 2010;
2. Dr. Qiang Cao, a visiting associate professor from China at UNL, who works on distributed file and storage systems with focus on scalability and reliability, partial support started in January 2009 and ended in June 2009.

(b) University of Maine – Orono (UNL), Award #: 0621493

Principal Investigator

- Dr. Yifeng Zhu

Graduate Research Assistants:

1. Cai Zhao, 3rd year Ph.D. student and Chinese citizen, working on adaptive metadata cache coherence protocols, support started in Jan. 2007
2. Ali Shareef, 2nd year Ph.D. student and U.S. Citizen, working on energy consumption modeling, partially supported by this project.
3. Lin Lin, 2nd year Ph.D. student at UMaine and Chinese citizen, working on metadata prefetching, partial support started in August 2008.
4. Jianhui Yue, 4th year Ph.D. student at UMaine and Chinese citizen, working on buffer caching, partial support started in August 2008.

(c) University of Central Florida (UCF), Subaward of 0621526

Principal Investigator

- Dr. Jun Wang

Graduate Research Assistants:

1. Peng Gu, defended his PhD dissertation in Summer 2008.
2. Ping Ge, Beginning Spring 2007, a 3rd PhD student, working on workload analyses and forecasting, full-time support since 2008.
3. Huijun Zhu, Beginning Spring 2007, a 3rd PhD student, working on workload analyses and forecasting, changed to part-time in Spring 2009.
4. Several Ph.D. students from the CASS (Computer Architecture and Storage System) group have been supported in part working on the project under the supervision of the PI Wang. We recruited two international graduate students in Fall 2009.

1.2 Other Organizations

1. Drs Rob Ross and Rajeev Thakur, DOE Argonne National Laboratory
2. Dr. Dan Feng, Wuhan National Laboratory for Optoelectronics and Huazhong University of Science and Technology, China
3. We have been in contact with Fermilab to monitor the metadata traffic of their dCache installation, as well as with the local US CMS tier2 site dCache installation.
4. Dr. Yifeng Zhu has recently received the Academic Excellence Grant (AEG) from the SUN Microsystems, Inc. SUN is donating equipment, valued at $63,960, for eight fully-configured Sun Fire T1000 servers to supplement and strengthen the HECURA research.

3. Project Activities and Findings

3.1 Research and education Activities:

Research Activities:

1. We have initiated monitoring for the eventual creation of traces of dCache usage both here at UNL as well as at FNAL.
2. We are in the process of designing a more scalable metadata server for dCache at UNL and U-Maine.
3. We have been conducting research to design and evaluate a significantly enhanced metadata management scheme, called G-HBA, based on our previous work on HBA, in terms of scalability and adaptivity over existing schemes at UNL and U-Maine.
4. We have been working on the design and implementation of an improved metadata prefetching scheme that is based on a data-mining technique at UNL and U-Maine.
5. We have been researching efficient schemes for metadata fault-tolerance and fault-recovery for significantly improved reliability and availability of metadata services at UNL.

6. We have been developing multi-variable forecasting models to analyze and predict file metadata access patterns at UCF.
7. We have been collecting cluster file system metadata access traces of some scientific benchmarks and applications at UCF.
8. We have been developing decentralized, locality-aware metadata grouping schemes Nexus grouping to facilitate the bulk metadata operations such as prefetching at UCF, UNL and U-Maine.
9. We have been developing scalable and adaptive Bloom filter arrays to enforce load balance and increase scalability for file mapping at UNL, U-Maine and UCF.

Education Activities:

1. The topic related to parallel I/O has been developed and taught in a new graduate-level course ECE 574 Cluster Computing at UMaine.
2. The topics related to reliability and availability of storage systems have been developed and incorporated in four senior/graduate level courses at UNL: CSCE 430/830 Computer Architecture, CSCE 432/832 High-Performance Processor Architecture, CSCE 488 Computer Engineering Professional Development, and CSCE 489 Computer Engineering Senior Design Project.
3. A research seminar course focusing on the scalability of distributed file systems in general and the Hadoop file system in particular, attended by both graduate and undergraduate students, has been developed and taught at UNL.

3.2 Findings:

1. RAID-6 significantly outperforms the other RAID levels in disk-failure tolerance due to its ability to tolerate arbitrary two concurrent disk failures in a disk array. The underlying parity array codes have a significant impact on RAID-6's performance. We propose a new XOR-based RAID-6 code, called the Partition Code (P-Code). P-Code is a very simple and flexible vertical code, making it easy to understand and implement. It works on a group of (prime-1) or (prime) disks, and its coding scheme is based on an equal partition of a specified two-integer-tuple set. P-Code has the following properties: (1) it is a Maximum-Distance-Separable (MDS) code, with optimal storage efficiency; (2) it has optimal construction and reconstruction computational complexity; (3) it has optimal update complexity (i.e., the number of parity blocks affected by a single data-block update is minimal). These optimal properties of P-Code are proven mathematically. While X-Code is provably optimal and RDP is proven optimal in computational complexity and storage efficiency, the latter in its current form is not optimal in update complexity. We propose a row parity placement strategy for RDP to help it attain optimal update complexity. P-Code

complements the other two optimal RAID-6 codes, X-code and RDP, to provide a near-full set of optimal RAID-6 configurations of typical disk-array size (e.g., 4-20 disks). That is, for any prime in a typical array size range, P-Code can be deployed for (prime-1) disks optimally, while X-code (or P-Code) and RDP can be respectively deployed for (prime) and (prime+1) disks optimally. Moreover, P-Code's potentially beneficial properties such as the flexible association between the blocks and their labels may find useful applications in distributed environments.

2. User I/O intensity has a significant impact on on-line RAID reconstruction performance by virtue of disk bandwidth contention. Based on this observation, we designed a novel scheme, called WorkOut, to significantly boost RAID reconstruction performance by I/O Workload Outsourcing. WorkOut effectively outsources all write requests and popular read requests originally targeted at the degraded RAID set to a surrogate RAID set during reconstruction. Our lightweight prototype implementation of WorkOut and extensive trace-driven and benchmark-driven experiments demonstrate that, compared with the existing reconstruction approaches, WorkOut significantly speeds up the reconstruction time and average user response time simultaneously. Importantly, WorkOut is orthogonal to and can be easily incorporated into any existing reconstruction algorithms. Furthermore, it is applicable to improving the performance of other background support RAID tasks such as resynchronization and disk scrubbing.

3. One of the challenging issues in performance evaluation of parallel storage systems through synthetic-trace-driven simulation is to accurately characterize the I/O demands of data-intensive scientific applications. We analyze several I/O traces collected from different distributed systems and conclude that correlations in parallel I/O inter-arrival times are inconsistent, either with little correlation or with evident and abundant correlations. Thus conventional Poisson or Markov arrival processes are inappropriate to model I/O arrivals in some applications. Instead, a new and generic model based on the α-stable process is proposed and validated to accurately model parallel I/O burstiness in both workloads with little and strong correlations. This model can be used to generate reliable synthetic I/O sequences in simulation studies. Experimental results show that this model can capture the complex I/O behaviors of real storage systems more accurately and faithfully than conventional models, particularly for the burstiness characteristics in the parallel I/O workloads.

4. A large-scale distributed file system must provide a fast and scalable metadata lookup service. In large-scale storage systems, multiple metadata servers are desirable for improving scalability. To this end we have proposed a novel scheme, called Group-

based Hierarchical Bloom Filter Array (G-HBA), judiciously utilizes Bloom filters to efficiently route requests to target metadata servers. Our G-HBA scheme extends the current Bloom filter-based architecture by considering dynamic and self-adaptive characteristics in ultra large-scale file systems. Our scheme logically organizes metadata servers (MDS) into a multi-layered query hierarchy and exploits grouped Bloom filters to efficiently route metadata requests to desired MDSs through the hierarchy. This metadata lookup scheme can be executed at the network or memory speed, without being bounded by the performance of slow disks. Experimental results show that this scheme can significantly improve metadata management scalability and query efficiency in ultra large-scale storage systems.

5. Obtaining representative and concise I/O workloads for the purpose of evaluating the performance of storage systems remains a challenge due to the complex nature of I/O behaviors. Previous studies have shown that disk I/O traffic can be represented as an independent and identically distributed random process in some workloads and a self-similar process in others. Additionally, workloads in the presence of self-similarity can exhibit either Gaussian or non-Gaussian characteristics. We have proposed a new and generic model based on the α-stable process to accurately build a synthetic workload representative of I/O traffic in production storage systems. The novelty of this new model is that it has the capability of characterizing both self-similar Gaussian and non-Gaussian workloads. Experimental results presented show that this model can more accurately capture the complex I/O behaviors of real storage systems than conventional models, particularly the burstiness and heavy-tail distribution under the Gaussian and non-Gaussian workloads.

6. File correlation, which refers to a relationship among related files that can manifest in the form of their common access locality (temporal and/or spatial), has become an increasingly important consideration for performance enhancement in Peta-scale storage systems. Previous studies on file correlations mainly concern with two aspects of files: file access sequence and semantic attribute. Based on mining with regard to these two aspects of file systems, various strategies have been proposed to optimize the overall system performance. Unfortunately, all of these studies consider either file access sequences or semantic attribute information separately and in isolation, thus unable to accurately and effectively mine file correlations, especially in large-scale distributed storage systems. We have developed a novel File Access corRelation Mining and Evaluation Reference model (FARMER) for optimizing Peta-scale file system performance that judiciously considers both file access sequences and semantic attributes simultaneously to evaluate the degree of file correlations by leveraging the Vector Space Model (VSM) technique adopted from the Information

Retrieval field. We extract the file correlation knowledge from some typical file system traces using FARMER, and incorporate FARMER into a real large-scale object-based storage system as a case study to dynamically infer file correlations and evaluate the benefits and costs of a FARMER-enabled prefetching algorithm for the metadata servers under real file system workloads. Experimental results show that FARMER can mine and evaluate file correlations more accurately and effectively. More significantly, the FARMER-enabled prefetching algorithm is shown to reduce the metadata operation latency by approximately 24-35% when compared to a state-of-the-art metadata prefetching algorithm and a commonly used replacement policy.

7. We also propose a simple but powerful on-line availability upgrade mechanism, Supplementary Parity Augmentations (SPA), to address the availability issue for parity-based RAID systems. The basic idea of SPA is to store and update the supplementary parity units on one or a few newly augmented spare disks for on-line RAID systems in the operational mode, thus achieving the goals of improving the reconstruction performance while tolerating multiple disk failures and latent sector errors simultaneously. By applying the exclusive OR operations appropriately among supplementary parity, full parity and data units, SPA can reconstruct the data on the failed disks with a fraction of the original overhead that is proportional to the supplementary parity coverage, thus significantly reducing the overhead of data regeneration and decreasing recovery time in parity-based RAID systems. In particular, SPA has two supplementary-parity coverage orientations, SPA Vertical and SPA Diagonal, which cater to user's different availability needs. The former, which calculates the supplementary parity of a fixed subset of the disks, can tolerate more disk failures and sector errors; whereas, the latter shifts the coverage of supplementary parity by one disk for each stripe to balance the workload and thus maximize the performance of reconstruction during recovery. The SPA with a single supplementary-parity disk can be viewed as a variant of but significantly different from the RAID5+0 architecture in that the former can easily and dynamically upgrade a RAID5 system to a RAID5+0-like system without any change to the data layout of the RAID5 system. Our extensive trace-driven simulation study shows that both SPA orientations can significantly improve the reconstruction performance of the RAID5 system while SPA Diagonal significantly improves the reconstruction performance of RAID5+0, at an acceptable performance overhead imposed in the operational mode. Moreover, our reliability analytical modeling and Sequential Monte-Carlo simulation demonstrate that both SPA orientations consistently more than double the MTTDL of the RAID5 system and improve the reliability of the RAID5+0 system noticeably.

8. To improve energy-efficiency, we propose RoLo (Rotated Logging), a novel logging architecture for write-intensive high-end computing that combines rotated logging with decentralized destaging to construct a logical logger with unlimited capacity. By rotating the active logging space among multiple disks, RoLo can effectively decrease energy consumption and increase reliability of replication-based storage systems, while avoiding the potential performance bottleneck and single point of failure of conventional logging architecture with extra dedicated log disks. We develop three flavors of RoLo, RoLo-E/R/P, to emphasize on energy efficiency, reliability, and performance respectively. Extensive trace-driven evaluations show their respective advantages. RoLo-E achieves up to 81.7% energy saving from a typical RAID10 system. RoLo-P provides the best performance among the RoLo schemes in most cases while RoLo-R achieves the highest reliability in terms of combined measure of MTTDL and disk-spin frequency at a performance cost of 3.77%-4.35% and no cost in energy efficiency.

9. We present DEBAR, a scalable and high-performance de-duplication storage system for backup and archiving, to overcome the throughput and scalability limitations of the state-of-the-art data de-duplication schemes, including the Data Domain De-duplication File System (DDFS). DEBAR uses a Two-Phase De-duplication Scheme (TPDS) that exploits memory cache and disk index properties to judiciously turn the notoriously random and small disk I/Os of fingerprint lookups and updates into large sequential disk I/Os, hence achieving a very high de-duplication throughput. The salient feature of this approach is that both the system backup and archiving capacity and the de-duplication performance can be dynamically and cost-effectively scaled up on demand; it hence not only significantly improves the throughput of a single de-duplication server but also is conducive to distributed implementation and thus applicable to large-scale and distributed storage systems.

10. We develop DSFS, a decentralized security system for large parallel file system. DSFS stores global access control lists (ACLs) in a centralized decision-making server and pushes pre-authorization lists (PALs) into storage devices. Thus DSFS allows users to flexibly set any access control policy for the global ACL or even change the global ACL system without having to upgrade the security code in their storage devices. With pre-authorization lists, DSFS enables a network-attached storage device to immediately authorize I/O, instead of demanding a client to acquire an authorization from a centralized authorization server at a crucial time. The client needs to acquire only an identity key from an authentication server to access any devices she wants. Experimental results show that DSFS achieves higher performance and scalability than traditional capability-based security protocols.

11. Prefetching is an effective technique for improving file access performance, which can significantly reduce access latency for I/O systems. In distributed storage systems, prefetching for metadata files is critical for the overall system performance. We have proposed an Affinity-based Metadata Prefetching (AMP) scheme for metadata servers in large-scale distributed storage systems to provide aggressive metadata prefetching. Through mining useful information about metadata accesses from past history, AMP can discover metadata file affinities accurately and intelligently for prefetching. Compared with LRU and some of the latest file prefetching algorithms such as Nexus and C-Miner, our trace-driven simulations show that AMP can improve buffer cache hit rates by up to 12%, 4.5% and 4% respectively, while reduce the average response time by up to 60%, 12% and 8%, respectively.

12. We develop SAFE, a Semantic- Aware source de-duplication Framework for Efficient cloud backup and restore. SAFE consists of three salient features, (1) Hybrid De-duplication, combining the global file- level de-duplication and local chunk-Level de-duplication to achieve an optimal tradeoff between the de-duplication efficiency and de-duplication overhead to achieve a short backup time; (2)Semantic-aware Elimination, exploiting the file semantics to narrow the search space for the duplicate files and data chunks in the hybrid de-duplication process; and (3)Unmodified Data Removal, removing the files and data chunks that are kept intact from data transmission for restore operations. Among these features, Hybrid De-duplication and Semantic-aware Elimination work in synch to remove the redundant data from data transmission to reduce backup times and storage costs, while Unmodified Data Removal aims to reduce the restore times. Through extensive experiments driven by real-world datasets, the SAFE framework is shown to maintain a much higher de-duplication efficiency/overhead ratio than existing solutions, shortening the backup times of the existing solutions by an average of 38.7%, and reduce the restore times by a ratio of up to 9.7 : 1.

13. We develop a novel decentralized semantic-aware metadata organization, called SmartStore, which exploits semantics of files' metadata to judiciously aggregate correlated files into semanticaware groups by using information retrieval tools. The key idea of SmartStore is to limit the search scope of a complex metadata query to a single or a minimal number of semantically correlated groups and avoid or alleviate brute-force search in the entire system. The decentralized design of SmartStore can improve system scalability and reduce query latency for complex queries (including range and top-k queries). Moreover, it is also conducive to constructing semantic-aware caching, and conventional filename-based point query. We have implemented a

prototype of SmartStore and extensive experiments based on real-world traces show that SmartStore significantly improves system scalability and reduces query latency over basic database approaches by more than one thousand times. To the best of our knowledge, this is the first study on the implementation of complex queries in large-scale file systems.

14. We develop a novel semantic-sensitive namespace management scheme, called Rapport, to provide dynamic and adaptive namespace management and support multiple queries with a constant-scale computational complexity. The basic idea behind Rapport is to build files' namespace by utilizing their semantic correlation and exploring the dynamic evolution of multi-dimensional attributes to support scalable and adaptive namespace management. The benefits of this approach are twofold. While the users receive the queried results from flexible queries, the file systems obtain significant performance improvements, especially in terms of scalability. Our extensive trace-driven experiments validate the effectiveness and efficiency of our proposed schemes. To the best of our knowledge, this is the first work on semantic-sensitive namespace management for ultra-scale file systems.

15. We develop MAD2, a scalable high-throughput exact deduplication approach for network backup services. MAD2 eliminates duplicate data both at the file level and at the chunk level by employing four techniques to accelerate the deduplication process and evenly distribute data. First, *M*AD2 organizes fingerprints into a Hash Bucket *M*atrix (HBM), whose rows can be used to preserve the data locality in backups. Second, M*A*D2 uses Bloom Filter *A*rray (BFA) as a quick index to quickly identify non-duplicate incoming data objects or indicate where to find a possible duplicate. Third, *D*ual Cache is integrated in MA*D*2 to effectively capture and exploit data locality. Finally, MAD*2* employs a *D*HT-based Load-Balance technique to evenly distribute data objects among multiple storage nodes in their backup sequences to further enhance performance with a well-balanced load. We evaluate our MAD2 approach on the backend storage of B-Cloud, a research-oriented distributed system that provides network backup services. Experimental results show that MAD2 significantly outperforms the state-of-the-art approximate deduplication approaches in terms of deduplication efficiency, supporting a deduplication throughput of at least 100MB/s for each storage component.

16. We develop a hybrid disk array architecture that combines a group of SSDs and two hard disk drives (HDDs) to improve the performance and reliability of SSD-based storage systems. In the proposed hybrid SSD-HHD disk array, the SSDs (data disks)

and part of one HDD (parity disk) compose a RAID4 disk array. Meanwhile, a second HDD and the free space of the parity disk are mirrored to form a RAID1-style write buffer that temporally absorbs the small write requests and acts as a surrogate set during recovery when a disk fails. The write data is reclaimed back to the data disks during the lightly loaded or idle periods of the system. Reliability analysis shows that the reliability of the hybrid disk array, in terms of MTTDL (Mean Time To Data Loss), is better than that of either HDD-based or SSD-based disk array. Our prototype implementation of the hybrid disk array and performance evaluations show that the proposed hybrid disk array significantly outperforms either HDD-based or SSD-based disk array

17. For parallel I/O workloads, the memory energy efficiency is determined by a complex interaction among four important factors: (1) cache hit rates that may directly translate performance gain into energy saving, (2) cache populating schemes that perform buffer allocation and affect access locality at the chip level, (3) request clustering that aims to temporally align memory transfers from different buses into the same memory chips, and (4) access patterns in workloads that affect the first three factors.We have developed a new energy-aware buffer cache replacement algorithm. Simulation results based on three real-world I/O traces, including TPC-R, web search and Finance applications, show that our algorithms can save up to 65.1% energy with marginal degradation in hit rates.

18. Indirect blocks, part of a file's metadata used for locating this file's data blocks, are typically treated indistinguishably from file's data blocks in buffer cache. We have found that this conventional approach significantly detriments the overall energy efficiency of memory systems. Scattering small but frequently accessed indirected blocks over all memory chips reduce the energy saving opportunities. We propose a new energy-efficient buffer cache management scheme, named MEEP, which separates indirect and data blocks into different memory chips. Our trace-driven simulation results show that our new scheme can save memory energy up to 20.8% for I/O-intensive server workloads.

19. To achieve a reasonably good trade-off among the three important storage design objectives of performance, reliability and energy-efficiency, we have developed an energy efficient disk array architecture, called a Green RAID (or GRAID), which extends the data mirroring redundancy of RAID10 by incorporating a dedicated log disk. The goal of GRAID is to significantly improve energy efficiency or reliability of existing RAID-based systems without noticeably sacrificing their reliability or energy efficiency. Reliability analysis shows that the reliability of GRAID, in terms of

MTTDL (Mean Time To Data Loss), is only slightly worse than RAID10 but much better than other existing RAID-based energy optimizing schemes such as EERAID and PARAID by up to 347 times, with an average of 155 times. On the other hand, our prototype implementation of GRAID and indirect comparisons show that GRAID's energy efficiency is significantly better than that of RAID10 by up to 32.1%, with an average of 25.4% while slightly better than or comparable to EERAID and PARAID.

20. Prefetching simply cannot help when the system is overloaded. We studied the scalability of the metadata servers equipped with Nexus prefetching algorithm by simulating large numbers of clients and servers. In the 16-server case, the throughput increases approximately 6% when the number of clients increase from 1000 to 2000, after that it stops growing since the system became saturated. With 64 or 256 servers, the system throughput scales up almost proportionally with the number of clients, indicating near optimal scalability of the system.

21. A lightweight segment structured local file system component named LSFS can be used to boost the local file metadata I/O performance for state-of-the-art parallel file systems. Parallel virtual file system (PVFS2) is chosen as an example for study. LSFS bridges the mapping gap by introducing a novel compact segment I/O technique, which facilitates the large-only raw disk I/O operations with the help of appropriate dynamic grouping algorithms. The current experimental results indicate that an LSFS-enhanced PVFS2 prototype system can significantly outperform a Linux-Ext3-based PVFS2 by up to 130% higher I/O bandwidth.

22. Recent years have seen a fast-growing volume of I/O traffic propagated through the local I/O interconnect bus. This raises up a question for storage servers on how to resolve such a potential bottleneck. We develop a hierarchical Data Cache Architecture called DCA to effectively slash local interconnect traffic and thus boost the storage server performance. A popular iSCSI storage server architecture is chosen as an example. DCA is composed of a read cache in NIC called NIC cache and a read/write unified cache in host memory called Helper cache. The NIC cache services most portions of read requests without fetching data via the PCI bus, while the Helper cache 1) supplies some portions of read requests per partial NIC cache hit, 2) directs cache placement for NIC cache, and 3) absorbs most transient writes locally. We develop a novel State-Locality-Aware cache Placement algorithm called SLAP to improve the NIC cache hit ratio for mixed read and write workloads. To demonstrate the effectiveness of DCA, we develop a DCA prototype system and evaluate it with an open source iSCSI implementation under representative storage server workloads.

Experimental results showed that DCA can boost iSCSI storage server throughput by up to 121 percent and reduce the PCI traffic by up to 74 percent compared with an iSCSI target without DCA.

23. Most of existing search algorithms for unstructured peer-to-peer (P2P) systems share one common approach: the requesting node sends out a keyword search query and the query message is repeatedly routed and forwarded to other peers in the overlay network. Due to multiple hops involved in query forwarding, the search may result in a long delay before it is answered. Furthermore, some incapable nodes may be overloaded when the query traffic becomes intensive or bursty. In this work, we develop a novel content-pushing, Advertisement-based Search Algorithm for unstructured Peer-to-peer systems (ASAP). An *advertisement* (ad) is a synopsis of contents a peer tends to share, and appropriately distributed and selectively cached by other peers in the system. In ASAP, nodes proactively advertise their contents by delivering ads, and selectively storing interesting ads received from other peers. Upon a request, a node can locate the destination nodes by looking up its local ads repository, and thus obtain a one-hop search latency with modest search cost. Comprehensive experimental results show that, compared with traditional query-based search algorithms, ASAP achieves much better search efficiency, and maintains system load at a low level with small variations. In addition, ASAP works well under node churn.

24. Abstract—Power consumption is an increasingly impressing concern for data servers as it directly affects running costs and system reliability. Prior studies have shown most memory space on data servers are used for buffer caching and thus cache replacement becomes critical. Temporally concentrating memory accesses to a smaller set of memory chips increases the chances of free riding through DMA overlapping and also enlarges the opportunities for other ranks to power down. This paper proposes a power and thermal-aware buffer cache replacement algorithm. It conjectures that the memory rank that holds the most amount of cold blocks are very likely to be accessed in the near future. Choosing the victim block from this rank can help reduce the number of memory ranks that are active simultaneously. We use three real-world I/O server traces, including TPC-C, LM-TBF and MSN-BEFS to evaluate our algorithm. Experimental results show that our algorithm can save up to 27% energy than LRU and reduce the temperature of memory up to 5.45oC with little or no performance degradation.

25. A challenging issue in performance evaluation of parallel storage systems through trace-driven simulation is to accurately characterize and emulate I/O behaviors in real

applications. The correlation study of inter-arrival times between I/O requests, with an emphasis on I/O intensive scientific applications, shows the necessity to further study the self-similarity of parallel I/O arrivals. This paper analyzes several I/O traces collected in large scale supercomputers and concludes that parallel I/Os exhibit statistically self-similar like behavior. Instead of Markov model, a new stochastic model is proposed and validated in this paper to accurately model parallel I/O burstiness. This model can be used to predicting I/O workloads in real systems and generate reliable synthetic I/O sequences in simulation studies.

4. Publications and Products

4.1 Journal Publications

1. Y. Zhu, H. Jiang, J. Wang and F. Xian, "HBA: Distributed Metadata Management System for Large Cluster-based Storage", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 19, No. 6, pp. 750-763, June 2008.
2. Y. Zhu, and H. Jiang, "RACE: A Robust Adaptive Caching Strategy for Buffer Cache", *IEEE Transactions on Computers*, Vol. 57, No. 1, pp. 25-40, January 2008
3. J. Yue, Y. Zhu and Z. Cai, "An Energy-Oriented Evaluation of Buffer Cache Algorithms Under Parallel I/O Workloads", *IEEE Transactions on Parallel and Distributed Systems*, Volume 19, Number 11, Pages 1565-1578, November 2008
4. Jun Wang, Peng Gu, Hailong Cai, An Advertisement-based Peer-to-Peer Search Scheme, preprint in *Journal of Parallel and Distributed Computing*.
5. Jun Wang, Xiaoyu Yao, Christopher Mitchell, and Peng Gu. A hierarchical data cache architecture for iSCSI storage server, *IEEE Transactions on Computers* Vol. 58, No. 4, pp 1-15, April 2009.
6. Xiao Qin and Hong Jiang. "Dynamic Load Balancing for I/O-Intensive Applications on Clusters," *ACM Transactions on Storage*, Vol. 5, No. 3, pp. 9:1-9:38, November 2009.
7. Jun Wang, Peng Gu, Hailong Cai. An Advertisement-based Peer-to-Peer Search Scheme, *Journal of Parallel and Distributed Computing*, Volume 69, Issue 7, July 2009, Pages 638-651.
8. Peng Gu, Jun Wang, Yifeng Zhu, Hong Jiang, Pengju Shang "A Novel Weighted-Graph-Based Grouping Algorithm for Metadata Prefetching," *IEEE Transactions on Computers,* Vol. 59, No. 1, pp. 1-15, January 2010
9. Xiao Qin, Hong Jiang, Adam Manzanares, Xiaojun Ruan, Shu Yin, "Communication-Aware Load Balancing for Parallel Applications on Clusters," *IEEE Transactions on Computers,* Vol. 59, No. 1, pp. 42-52, January 2010.

10. Yu Hua, Yifeng Zhu, Hong Jiang, Dan Feng, and Lei Tian, "Supporting Scalable and Adaptive Metadata Management in Ultra Large-scale File Systems," *IEEE Transactions on Parallel and Distributed Systems.* (2010) [Accepted]
11. X. Qin, H. Jiang, A. Manzanares, X.-J Ruan, and S. Yin. "Dynamic Load Balancing Support for I/O-Intensive Jobs in Homogeneous and Heterogeneous Clusters of Workstations." *IEEE Transactions on Computers.* (2010). [Accepted]
12. Saba Sehrish and Jun Wang, "Reduced Function Set Architecture for MPI-IO," accepted by *Journal of Supercomputing*.

4.2 Book(s) of other one-time publications(s):

<u>*Paper Under Review:*</u>

1. Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, and Lei Tian, "Semantic-Aware Metadata Organization Paradigm in Next-Generation File Systems", *IEEE Transactions on Parallel and Distributed Systems.*
2. Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao, "Improving Availability of RAID-Structured Storage Systems by Workload Outsourcing", *IEEE Transactions on Computers.*
3. Lei Tian, Hong Jiang, Dan Feng, Qiang Cao, Changsheng Xie, and Qin Xin, "SPA: On-Line Availability Upgrades for Parity-based RAIDs through Supplementary Parity Augmentations," *ACM Transactions on Storage*.
4. Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, "Rapport: Semantic-sensitive Namespace Management in Large-scale File Systems," submitted to *The 23rd International Conference on High Performance Computing, Networking, Storage and Analysis (The 23rd Annual Supercomputing Conference -- SC'10)*, New Orleans, Lousiana, November 13-19, 2010.
5. Zhichao Yan, Hong Jiang, Dan Feng, Lei Tian, Yujuan Tan, and Jingning Liu, "SUV-TM: A Novel Single-Update Version-Management Scheme for Hardware Transactional Memory Systems," submitted to *The 16th International Conference on High Performance Computing (HiPC'10),* Goa, India, December 19-22, 2010.
6. Zhongying Niu, Hong Jiang, Ke Zhou, Dan Feng, "DSFS: Decentralized Security for Large Parallel File Systems", submitted to *The 11th ACM/IEEE International Conference on Grid Computing (Grid 2010),* Brussels, Belgium, Oct 24 – Oct 29, 2010.
7. Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, Zhichao Yan, and Guohui Zhou, "SAFE: A Semantic-Aware Source De-duplication Framework for Efficient cloud backup and restore," submitted to *Software: Practice and Experience.*

<u>*Paper Published:*</u>

1. Jian Hu, Hong Jiang, Lei Tian, Lei Xu, "PUD-LRU: An Erase-Efficient Write Buffer Management Algorithm for Flash Memory SSD," accepted to appear in the *Proceedings of The 18th Annual Meeting of the IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS10)*, Miami, Florida, August 17-19, 2010.
2. Yujuan Tan, Hong Jiang, Dan Feng, Lei Tian, and Zhichao Yan, "SAM: A Semantic-Aware Multi-Tiered Source De-duplication Framework for Cloud Backup," accepted to appear in the *Proceedings of The 39th International Conference on Parallel Processing (ICPP 2010)*, San Diego, CA, September 13-16, 2010.
3. Yang Hu, Hong Jiang, Dan Feng, Lei Tian, Sshuping Zhang, Jingning Liu, Wei Tong, "Achieving Page-Mapping FTL Performance at Block-Mapping FTL Cost by Hiding Address Translation," *Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010),* Incline Village, Nevada, May 3-7, 2010.
4. Jiansheng Wei, Hong Jiang, Ke Zhou, Dan Feng,"MAD2: A Scalable High-Throughput Exact Deduplication Approach for Network Backup Services," *Proceedings of The 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST2010),* Incline Village, Nevada, May 3-7, 2010.
5. Yinliang Yue, Hong Jiang, Lei Tian, Fang Wang, Dan Fang, and Quan Zhang. "RoLo: A Rotated Logging Storage Architecture for Enterprise Data Centers," *Proceedings of The 30th International Conference on Distributed Computing Systems (ICDCS 2010)*, Genoa, Italy, June 21-25, 2010.
6. Dongyuan Zhan, Hong Jiang, and Sharad Seth, "Exploiting Set-Level Non-Uniformity of Capacity Demand to Enhance CMP Cooperative Caching", *Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010)*, Atlanta, GA, April 19-23, 2010.
7. Tianming Yang, Hong Jiang, Dan Feng, Zhongying Niu,Ke Zhou, andYaping Wan, "DEBAR: A Scalable High-Performance De-duplication Storage System for Backup and Archiving," *Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010)*, Atlanta, GA, April 19-23, 2010.
8. Bo Mao, Hong Jiang, Dan Feng, Suzhen Wu, "HPDA: A Hybrid Parity-based Disk Array for Enhanced Performance and Reliability," *Proceedings of the 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010)*, Atlanta, GA, April 19-23, 2010.
9. J. Yue, Y. Zhu, Z. Cai, L. Lin, "Energy and Thermal Aware Buffer Cache Replacement Algorithm", in Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2010 (Acceptance Rate: 18/53 = 34%)
10. Q. Zou, Y. Zhu and D. Feng, "A study of Self-similarity in Parallel I/O Workloads", in Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), 2010 (short paper)

11. Suzhen Wu, Dan Feng, Hong Jiang, Bo Mao, Lingfang Zeng, and Jianxi Chen, “JOR: A Journal-guided Reconstruction Optimization for RAID-Structured Storage Systems", *Proceedings of the Fifteenth International Conference on Parallel and Distributed Systems (ICPADS'09),* Shenzhen, China, December 8-10, 2009.
12. Hui Tian, Ke Zhou, Hong Jiang, Dan Feng, “Digital Logic Based Encoding Strategies for Voice-over-IP Steganography,” *Proceedings of the ACM Multimedia 2009 conference (ACM-MM’09*), Beijing, China, October 19-24, 2009.
13. Yu Hua, Hong Jiang, Yifeng Zhu, Dan Feng, and Lei Tian. "SmartStore: A New Metadata Organization Paradigm with Metadata Semantic-Awareness for Next-Generation File Systems." *Proceedings of The 22nd International Conference on High Performance Computing, Networking, Storage and Analysis (The 22nd Annual Supercomputing Conference -- SC'09)*, Portland, Oregon, November 14-20, 2009.
14. Saba Sehrish*, Jun Wang, and Rajeev Thakur (ANL). Self-detecting Locks to Support MPI-IO Atomicity. EuroPVM/MPI September 2009. (8 pages)
15. Saba Sehrish, and Jun Wang. "Smart Read/Write for MPI-IO". In *the 14th International Workshop on High-Level Parallel Programming Models and Supportive Environments, in conjunction with the 23rd IEEE International Parallel and Distributed Processing Symposium*. May 2009.
16. Chao Jin, Hong Jiang, Dan Feng, Lei Tian, "P-Code: A New RAID-6 Code with Optimal Properties," in the *Proceedings of the 23rd ACM International Conference on Supercomputing (ICS ’09)*, IBM T.J. Watson Research Center, New York, NY, June 8-12, 2009.
17. Hui Tian, Ke Zhou, Hong Jiang, Yongfeng Huang, Jin Liu, and Dan Feng, "An Adaptive Steganography Scheme for Voice over IP," accepted to appear in the *Proceedings of the 2009 IEEE International Symposium on Circuits and Systems (ISCAS’09)*, Taipai, Taiwan, May 24-27, 2009.
18. Hui Tian, Ke Zhou, Hong Jiang, Yongfeng Huang, Jin Liu, and Dan Feng, "An M-Sequence Based Steganography Model for Voice over IP," accepted to appear in the *Proceedings of the 2009 IEEE International Conference on Communications (ICC’09)*, Dresden, Germany, June 14-18, 2009.
19. Suzhen Wu, Hong Jiang, Dan Feng, Lei Tian, and Bo Mao, "WorkOut: I/O Workload Outsourcing for Boosting the RAID Reconstruction Performance, " in the *Proceedings of the 7th USENIX Conference on File and Storage Technologies (FAST '09)*, San Francisco, CA, USA, February 24-27, 2009, pp. 239-252.
20. Y. Hua, H. Jiang, Y. Zhu, D. Feng, L. Tian, "SmartStore: A New Metadata Organization Paradigm with Semantic-Awareness, " *7th USENIX Conference on File and Storage Technologies, Work-In-Progress*, Feb., 2009, San Francisco

21. Y. Zhu and H. Jiang, "Efficient Update Control of Bloom Filter Replicas in Large-Scale Distributed Systems", Book chapter*, Handbook of research on scalable computing technologies*, IGI Global, USA, 2009
22. D. Feng, Q. Zou, H. Jiang, and Y. Zhu, "A Novel Model for Synthesizing Parallel I/O Workloads in Scientific Applications," in the *Proceedings of IEEE International Conference on Cluster Computing*, pp. 252-261, Tsukuba, Japan, Sept. 29 - Oct. 1, 2008
23. Q. Zou, D. Feng, Y. Zhu, H. Jiang, X. Ge, and Z. Zhou, "A Novel and Generic Model for Synthesizing Disk I/O Traffic Based on The Alpha-stable Process," in the *Proceedings of 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '08)*, Baltimore, MD, Sept 8-10, 2008
24. Bo Mao, Dan Feng, Hong Jiang, Suzhen Wu, Jianxi Chen, Lingfang Zeng, "GRAID: A Green RAID Storage Architecture with Improved Energy Efficiency and Reliability, " in the *Proceedings of the 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '08)*, Baltimore, MD. USA, September 8-10, 2008.
25. J. Yue, Y. Zhu, Z. Cai, "Energy Efficient Buffer Cache Replacement", in the *Proceedings of 16th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '08)*, Baltimore, MD, Sept 8-10, 2008 (Poster Paper)
26. J. Yue, Y. Zhu, and Z. Cai, "Impacts of Indirect Blocks on Buffer Cache Energy Efficiency", in *Proceedings of the 37th International Conference on Parallel Processing (ICPP '08)*, Portland, Oregon, USA, September 2008
27. Y. Hua, Y. Zhu, H. Jiang, D. Feng, and L. Tian, "Scalable and Adaptive Metadata Management in Ultra Large-Scale File Systems", in the *Proceedings of the 28th International Conference on Distributed Computing Systems (ICDCS '08)*, June 17-20, 2008
28. L. Lin, M. Li, H. Jang, Y. Zhu and L. Tian, "AMP: An Affinity-based Metadata Prefetching Scheme in Large-Scale Distributed Storage Systems", in the *Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid '08)*, May 19-22, 2008, Lyon, France
29. Hailong Cai, Ping Ge, Jun Wang. "Applications of Bloom Filters in Peer-to-peer Systems: Issues and Questions". In the *Proceedings of International Conference on Networking, Architecture, and Storage (NAS '08)*.
30. Peng Gu, Jun Wang, Robert Ross, "Bridging the Gap Between Parallel File Systems and Local File Systems: A Case Study with PVFS," In the *Proceedings of ICPP* 2008.

31. J. Yue, Y. Zhu and Z. Cai, "Evaluating Memory Energy Efficiency in Parallel I/O Workloads", in the *Proceedings of IEEE International Conference on Cluster Computing*, September, 2007, Austin, TX, pp. 21-30 (Best Paper Award)
32. ASAP: An Advertisement-based Search Scheme for Unstructured Peer-to-Peer Systems. Peng Gu, Jun Wang, Hailong Cai. In the *Proceedings of International Conference on Parallel Processing (ICPP '07)*, Sept. 2007. Xi'an, China.
33. Lei Tian, Dan Feng, Hong Jiang, Ke Zhou, Lingfang Zeng, Jianxi Chen, Zhikun Wang, and Zhenlei Song, "PRO: A Popularity-based Multi-threaded Reconstruction Optimization for RAID-Structured Storage Systems", in the *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07)*, San Jose, CA, February 13-16, 2007.
34. Yifeng Zhu and Hong Jiang, "False Rate Analysis of Bloom Filter Replicas in Distributed Systems," in the *Proceedings of the 2006 International Conference on Parallel Processing (ICPP '06)*, Columbus, Ohio, August 14-18, 2006.

4.2 Internet Dissemination

- http://www.eece.maine.edu/research/sam2/
- http://t2.unl.edu/documentation/hecura
- http://www.eecs.ucf.edu/~jwang/research-metadata.html

5. Contributions

5.1 Contributions within Discipline

Scientific computing brings revolution to basic scientific and engineering research. Through simulation rather than physical experiments, new knowledge can be discovered and new technology can be developed more quickly, efficiently and cost-effectively in many areas, especially where experiments are expensive, hazardous, or even impossible to perform or observe. However, data I/O tends to be the performance bottleneck for many scientific simulations. This research project aims to eliminate one of the major bottlenecks in data I/O operations: metadata bottlenecks. The following highlights our research outcomes.

1. Hierarchical directory do not meet scalability and functionality requirements for the next-generation extra-large storage systems. We design a semantic-aware organization, called SmartStore, which exploits metadata semantics of files to limits metadata complex queries to a single or a minimal number of semantically related groups. Experimental results show that Smartstore improves system scalability and reduces query latency over basic database approaches by one thousand times (SC'2009).

2. In extra-large storage systems, users often need to manually navigating hierarchies of billions of files or directories to locate target data. We develop a new metadata management scheme, called Rapport, use semantic correlations, instead of file names, to represent a file. It makes use of semantic correlation to create a semantic-sensitive namespace, which consists of the most closely correlated files identified by using a simple and fast LSH-based computation. Experimental results show that Rapport is very scalable and efficient (submitted to SC'2010).
3. We exploit the semantic information to improve the storage systems. Particularly, we design two de-duplication frameworks, called SAFE and MAD2, for high-throughput backup and restore. SAFE combines the global file-level de-duplication and local chunk-level de-duplication to achieve an optimal tradeoff between the de-duplication efficiency and de-duplication overhead to achieve a short backup time. MAD2 eliminates duplicate data both at the file level and at the chunk level by employing four techniques to accelerate the de-duplication process and evenly distribute data. Experiments shows that MAD2 can achieve at least 100 MB/s for each storage component, and SAFE can shorten the backup times of the existing solutions by an average of 38.7%, and reduce the restore times by a ratio of up to 9.7 : 1. (MSST'2010).
4. A significant extension of our HBA scheme and its prototype implementation that were completed during the first year of the funding, G-HBA, for name space management of ultra large-scale file systems has been developed and prototyped as well (ICDCS'08). The G-HBA scheme is significantly more scalable, adaptive, and space-efficient than HBA;
5. A novel file correlation mining and evaluation model, called FARMER (HPDC'08), has been developed to accurately and efficiently explore and exploit file correlations to optimize storage system performances. To the best of our knowledge, this is the first such scheme that exploits the coordination and correlation between dynamic semantics of file popularity (access patterns) with static semantics of files to accurately mine file correlations.
6. A test framework has been set up to gather information about metadata access behaviors in dCache (implemented at UNL's tier-2 CMS site), and substantial redesign and streamlining have been done to the current dCache metadata management to enhance the scalability and reliability of the dCache metadata management. Preliminary results have been promising.
7. We propose a new energy-efficient buffer cache management scheme, named MEEP, which separates indirect and data blocks into different memory chips. Our trace-driven simulation results show that our new scheme can save memory energy up to 20.8% for I/O-intensive server workloads.

8. An extended Parallel Virtual File System prototype system with enhanced local I/O component has been developed and tested on several medium scaled cluster systems. Comprehensive experimental results indicate the system throughput can be boosted by up to 132% using several parallel I/O benchmarks.
9. A substantial fault-recovery mechanism, WorkOut is proposed and developed to boost the reconstruction performance for RAID systems (FAST'09). More importantly, WorkOut is orthogonal to and can be easily incorporated into any existing reconstruction algorithms. Furthermore, it is also applicable to improving the performance of other background support RAID tasks such as re-synchronization and disk scrubbing.
10. A novel RAID-6 coding scheme with optimal properties of optimal storage efficiency, construction and reconstruction computational complexity, and update complexity, P-Code is proposed to resist from double disk failures efficiently (ICS '09). P-Code is a very simple and flexible vertical code, making it easy to understand, implement, and deploy.
11. R-tree with Bloom Filters (RBF) is being developed. A New Storage Structure for Space-Efficient Queries for Multidimensional Metadata in OSS (WiP of FAST07). We propose a new space-efficient storage structure, called the R-tree with Bloom Filters (RBF), to store multidimensional metadata and achieve point and range query with low operational complexity. The basic idea of our RBF is to expand the classical R-tree to incorporate space-efficient Bloom filters in R-tree nodes, maintaining multidimensional range information and achieving space efficiency.
12. We are improving metadata reliability through popularity and locality based reconstruction schemes (work in progress; preliminary work on file data (read most) was recently presented at FAST'07);
13. Appropriate interdisciplinary use of statistical models is being identified to analyze the metadata access patterns of both local file systems and cluster-based file systems.

5.2 Contributions to Other Disciplines

1. The research project has started to integrate into the US CMS research facility at the University of Nebraska (UNL). US CMS is a collaboration US scientists participating in the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland. UNL's US CMS Tier-2 site is a child site of the Tier-1 site at Fermi Nation Laboratory (FNAL). CMS sites employ dCache, a distributed storage data caching system, to support data access and transfer. We have started to prototype SAM2 toolkit, in particular the prefetching algorithms, into dCache to improve the I/O performance. In addition to CMS Tier-2 facility, UNL's Research Computing

Facility (RCF), the primary computation resource at UNL, will benefit from our SAM^2 toolkit. RCF includes (1) a distributed-memory supercomputer named Prairiefire that has 256 AMD Opteron processors capable of 88.5Gflops and (2) a shared-memory supercomputer from SGI named Homestead that contains 32 500MHz MIPS processor.

2. At the University of Maine, Dr. Yifeng Zhu is working with researchers in Marine Sciences to alleviate the I/O bottleneck for their simulations. The ocean model developed at UMaine is I/O intensive. We are working to utilize parallel I/O to speed up their applications.

5.3 Contributions to Human Resources and Education

At the University of Central Florida, Peng Gu defended his Ph.D. dissertation in June 2008.

At the University of Maine, the research findings and concepts are being incorporated into two innovative NSF-funded education programs, leaded by Dr. Yifeng Zhu, to provide college undergraduates as well as middle-school teachers and their students' firsthand experiences in scientific computing. (1) The Supercomputing Undergraduate Program in Maine (SuperMe), funded by a $300,000 grant from NSF, is an opportunity for 10 UMaine undergraduate students to spend the summer conducting the kind of sophisticated, meaningful scientific research that is usually reserved for more advanced students. (2) With a separate $1.2 million NSF grant, another three-year program aims to integrate supercomputer modeling into the Maine middle-school science curriculum. Called Inquiry-based Dynamic Earth Applications of Supercomputing (IDEAS), the program will allow 20 middle-school teachers and 60 of their students each year to explore the myriad intricacies of UMaine's climate computer model by accessing the supercomputer with their state-issued laptops.

5.4 Contributions to Resources for Science and Technology

5.5 Contributions Beyond Science and Engineering