10-6-2014

DC: Small: Energy-aware Coordinated Caching in Cluster-based Storage Systems

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

Preview of Award 0916663 - Final Project Report

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	0916663
Project Title:	DC:Small: Energy-aware Coordinated Caching in Cluster-based Storage Systems
PD/PI Name:	Yifeng Zhu, Principal Investigator
Recipient Organization:	University of Maine
Project/Grant Period:	09/01/2009 - 12/31/2013
Reporting Period:	09/01/2013 - 12/31/2013
Submitting Official (if other than PD\PI):	Yifeng Zhu Principal Investigator
Submission Date:	10/06/2014
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Yifeng Zhu

Accomplishments

*** What are the major goals of the project?**

The main goal of this project is to improve the performance and energy efficiency of I/O operations of large-scale cluster computing platforms.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities: The major activities include: 1) characterize the memory access workloads; 2) investigate the new and emerging new storage and memory devices, such as SSD and PCM, on I/O performance. (3) study energy-efficient buffer and cache replacement algorithms, (4) leveraging SSD as a new caching device to improve the energy efficiency and performance of I/O performance

Specific Objectives: **(1) Analysis and synthesis of memory workloads and parallel I/O workload**

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. We have analyzed several I/O traces collected in large-scale supercomputers and concluded that parallel I/Os exhibit statistically self-similar like behavior.

(2) Energy efficient buffer cache and memory management

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. We have designed a new power and thermal-aware buffer cache replacement algorithm.

In addition, row accesses in memory chips are not only very slow in response but also cost significant amount of energy. The interleaved access from different process segments destroys access locality seen at process segment. To address this, we design a new memory architecture that adds a small cache in memory controller to recover accesses locality and a new cache management scheme that exploits the semantic information of memory access requests to better capture the access locality.

(3) Caching in hybrid storage systems

Buffer cache replacement schemes play an important role in conserving memory energy, since buffer cache is frequently more than 77% of the total available memory on desktop computers and even more on storage servers. Specifically, memory energy is impacted from two aspects: (1) Replacement algorithms with high hit rates help reduce the overall running time and thus save energy directly; (2) Replacement algorithms determine the access sequence and utilization of memory chips, and hence influence the opportunities of powering down and DMA overlapping, as explained in detail later. Most replacement algorithms aim only to maximize cache hit rates and ignore the current power status of memory chips when selecting a victim block upon a cache miss. As result, energy saved due to higher hit rates and shorter running time may not offset the extra energy cost when keeping more memory chips simultaneously active. This observation motivates us to study new cache replacement algorithms that optimize the tradeoff between cache hit rates and energy-efficiency.

Hybrid storage systems leverages the fast random access performance in SSDs to boost the overall I/O performance without generating a large cost overhead. The key challenging research issue in such a hybrid storage system is how to dynamically allocate or migrate data between the SSD and the disks in order to achieve the optimal performance gain. In this paper, we propose a hybrid storage architecture that treats the SSD as a by-passable cache to hard disks, and design an online algorithm that judiciously exchanges data between the SSD and the disks. Our basic principle is to place hot and randomly accessed data on the SSD, and other data, particularly cold and sequentially accessed data on hard disks. Our hybrid storage system, called Hot Random Off-loading (HRO), is implemented as a simple and general user-level layer above conventional file systems in Linux and supports standard POSTIX interfaces, thus requiring no modifications to underneath file systems or users

applications. This prototype is comprehensively evaluated by using a commodity hard disk and SSD.

(4) Leveraging emerging memory devices to improve the performance

Exascale computing will require 1000 times more memory than available today. However, today's DRAM technology is hitting the wall of energy efficiency and transistor scaling. It is a great challenge to fabricate high density DRAM beyond 22nm due to difficulties such as efficient charge placement and capacitor control, and reliable charge sensing. Energy consumption and heat dissipation of DRAM with large capacities is a severe issue.

Fortunately, emerging memory devices, such as phase-change memory (PCM) has better process scalability and less leakage power, providing a viable alternative to DRAM in the near future. However, one of the major weaknesses of PCM is slow write performance.

We proposed two new mechanisms (PASAK and WAVAK) that leverage subarray-level parallelism to enable a bank to serve a write and multiple reads in parallel without violating power constraints. We also designed a new coding scheme to improve the speed of the write-1 stage by further increasing the number of bits that can be written to PCM in parallel.

(5) Using memory buffer and SSD cache to improve the performance of shingled-recording disk systems

The areal density of magnetic disks is reaching its length-scale limitation. Shingled recording technology is the most promising one since it can notably increase the grain density without changing underline storage media. The key challenge of extending the application of shingled writing and integrating it into magnetic disk system is to lower the write amplification of shingled writing disk. A different approach to alleviate the random write issue is hybrid systems that leverage a small non-volatile RAM or SSD for effectively caching hot data and reducing data immigration and improve performance. Currently SSD outperform disks significantly in both read and write, particularly in random read. Using SSD for caching random reads is very helpful to improve the performance of shingled disk system.

Significant Results: **(1) Analysis and synthesis of memory workloads and parallel I/O workload**

We have analyzed several I/O traces collected in large-scale supercomputers and concluded 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. We have also studied the auto-correlation functions of the arrival intervals of memory accesses in all SPEC CPU2006 traces, and concluded that correlations in memory inter-access times are inconsistent, either with evident correlations or with little and no correlation. Different with the studies focused on the prior suites, we present that self-similarity exists only in a small number of SPEC2006 workloads. In addition, we implement a memory access series generator in which the inputs are the measured properties of the available trace data. Experimental results show that this model can more accurately emulate the complex access arrival behaviors

of real memory systems than the conventional self-similar and independent identically distributed methods, particularly the heavy-tail characteristics under both Gaussian and non-Gaussian workloads.

Publications:

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), pp. 1-6, 2010

Q. Zou, J. Yue, Y. Zhu, and B. Segee, "Temporal Characterization of SPECCPU2006 Workloads: Analysis and Synthesis", in Proceedings of the 31st IEEE International Performance Computing and Communications Conference (IPCCC), pp. 11-20, 2012

(2) Energy efficient buffer cache and memory management

We have designed a power and thermal-aware buffer cache replacement algorithm for data servers. 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.

In order to slow row access in memory chips, we have designed a new memory architecture that adds a small cache in memory controller to recover accesses locality and a new cache management scheme that exploits the semantic information of memory access requests to better capture the access locality. Specifically, the accesses to different segments of a process, such as code segment and data segment, are stored and managed separately in the cache. Our experiments show that our new scheme can achieve up to 87.5% DRAM energy reduction and up to 63.5% benchmark completion time reduction. In addition, we find that our design with sequential physical address mapping is superior to interleave physical address mapping in term of both DRAM energy and benchmark completion time, which is different from conventional understanding.

Publications:

J. Yue, Y. Zhu, Z. Cai, L. Linz, "Energy and Thermal Aware Buffer Cache Replacement Algorithm", in Proceedings of 26th IEEE Symposium on Massive Storage Systems and Technologies (MSST), pp. 1-10, 2010

J. Yue, Y. Zhu, Z. Cai, and L. Lin, "Energy Efficient Buffer Cache Replacement for Data Servers", in Proceedings of the 2011 International Conference on Networking, Architecture, and Storage (NAS), pp. 328-338, 2011

(3) Caching in hybrid storage systems

Random accesses are generally harmful to performance in hard disk drives due to more dramatic mechanical movement. This paper presents the design, implementation, and evaluation of Hot Random Off-loading (HRO), a self-optimizing hybrid storage system that uses a fast and small SSD as a by-passable cache to hard disks, with a goal to serve a majority of random I/O accesses from the fast SSD. HRO dynamically estimates the performance benefits based on history access patterns, especially the randomness and the hotness, of individual files, and then uses a 0-1 knapsack model to allocate or migrate files between the hard disks and the SSD. HRO can effectively identify files that are more frequently and randomly accessed and place these files on the SSD. We implement a prototype of HRO in Linux and our implementation is transparent to the rest of the storage stack, including applications and file systems. We evaluate its performance by directly replaying three real-world traces on our prototype. Experiments demonstrate that HRO improves the overall I/O throughput up to 39% and the latency up to 23%.

Publications:

L. Lin, Y. Zhu, J. Yue, Z. Cai and B. Segee, "Hot Random Off-loading: A Hybrid Storage System With Dynamic Data Migration", in Proceedings of the 19th Annual Meeting of the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOSTS), pp. 318-325, 2011

(4) Leveraging emerging memory devices to improve the performance

To improve the write performance of PCM, we propose a new write scheme, called two-stage-write, which leverages the speed and power difference between writing a zero bit and writing a one bit. Writing a one takes longer time but less electrical current than writing a zero. We propose to divide a write into stages: in the write-0 stage all zeros are written at an accelerated speed, and in the write-1 stage, all ones are written with increased parallelism, without violating power constraints. We also present a new coding scheme to improve the speed of the write-1 stage by further increasing the number of bits that can be written to PCM in parallel. Based on simulation experiments of a multi-core processor under various SPEC CPU 2006 workloads, our proposed techniques can reduce the memory latency of standard PCM by 68.3% and improve the system performance by 33.9% on average. In addition, we designed new write techniques that can better leverage subarray-level parallelism and fully overlap a write operation with read operations that access different subarrays within the same bank.

Publications:

J. Yue, Y. Zhu, "Making Write Less Blocking for Read Accesses in Phase Change Memory", in Proceedings of The 20th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), pp. 269-277, 2012

J. Yue, Y. Zhu, "Accelerating Write by Exploiting PCM Asymmetries", in Proceedings of the 19th IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 282-293, 2013, ***Best Paper Nominee***

J. Yue, Y. Zhu, "Exploiting Subarrays Inside a Bank to Improve Phase Change Memory Performance", in Proceedings of Design, Automation & Test in Europe (DATE), pp. 386-391, 2013

(5) Using memory buffer and SSD cache to improve the performance of shingled-recording disk systems

We design a hybrid wave-like shingled recording disk system (HWSR) to improve both the performance and the capacity of a shingled recording disk. HWSR contains three different storage media: memory, SSD, and hard disk. The memory has a very small capacity, such as 100MB, in our design to reduce the overall cost. It is used to buffer hot writes. The SSD is used as a disk cache to improve random read performance.

Publication:

J. Wan, N. Zhao, Y. Zhu, J. Wang, Y. Mao, and C. Xie, "High Performance and High Capacity Hybrid Shingled-Recording Disk System", in Proceedings of IEEE International Conference on Cluster Computing (CLUSTER), pp. 173-181, 2012

D. Luo, J. Wan, N. Zhao, Y. Zhu, P. Chen, N. Zhao, F. Li, and C. Xie, "Design and Implementation of A Hybrid Shingled Write Disk System", submitted to IEEE Transaction on Parallel and Distributed Computing (14 pages, under review)

Key outcomes or Other achievements: This project has published more than 10 research papers in top conferences and journals. It has also trained two Ph.D. students and 1 M.S. student at the University of Maine (UMaine).

*** What opportunities for training and professional development has the project provided?**

This project has partially trained two Ph.D. students and one M.S. student at the University of Maine.

*** How have the results been disseminated to communities of interest?**

We have published multiple papers in top conferences and journals, including DATE'13, HPCA'13, IEEE TPDS, Cluster'12, NAS'12, IPCCC'12, MASCOTS'12, and CCGrid'12. Our HPCA'13 paper was one of the four candidates for Best Paper Award.

In addition, all publications can be found from the project web site: http://www.eece.maine.edu/~zhu/storage

Products

Books

Book Chapters

Conference Papers and Presentations

Inventions

Journals

A. Shareef, Y. Zhu (2012). Effective Stochastic Modeling of Energy Constrained Wireless Sensor Networks. *Special issue of Energy-Efficient Wireless Communications with Future Networks and Diverse Devices (EEWC), Journal of Compute Networks and Communications*. 2012 20 pages. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

M. Yu, J. Wan, Y. Zhu, C. Xie (). A New Parity-Based Migration Method to Expand RAID-5. *IEEE Transactions on Parallel and Distributed Systems*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Y. Hua, J. Hong, Y. Zhu, D. Feng, X. Lei (2014). SANE: Semantic-Aware Namespace in Ultra-large-scale File Systems. *IEEE Transactions on Parallel and Distributed Systems*. 25 (5), 1328-1338. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Licenses

Other Products

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations

Websites
Project website
http://www.eece.maine.edu/~zhu/storage

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Zhu, Yifeng	PD/PI	1
Yue, Jianhui	Postdoctoral (scholar, fellow or other postdoctoral position)	6
Cai, Zhao	Graduate Student (research assistant)	6
Maskay, Aman	Undergraduate Student	2

Full details of individuals who have worked on the project:

Yifeng Zhu
Email: zhu@eece.maine.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Supervising and advising graduate students; Leading the projects; Integrating research into teaching.

Funding Support: None

International Collaboration: Yes, China
International Travel: Yes, China - 0 years, 0 months, 14 days; France - 0 years, 0 months, 7 days

Jianhui Yue
Email: jianhui.yue@maine.edu
Most Senior Project Role: Postdoctoral (scholar, fellow or other postdoctoral position)
Nearest Person Month Worked: 6

Contribution to the Project: Dr. Yue mainly worked on leveraging emerging new memory devices to improve the energy efficiency and performance of I/O accesses.

Funding Support: Dr. Yue is partially supported by this grant.

International Collaboration: No
International Travel: No

Zhao Cai
Email: cai.zhao@maine.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 6

Contribution to the Project: Mr. Cai mainly worked on improve the performance and energy efficiency of disk I/O devices by designing new journal file systems and using SSD.

Funding Support: Mr. Cai is supported by this grant.

international Collaboration: No
international Travel: No

Aman Maskay
Email: aman.maskay@maine.edu
Most Senior Project Role: Undergraduate Student
Nearest Person Month Worked: 2

Contribution to the Project: He investigated different embedded computer systems boards that can be used to support hybrid storage devices, which combines a SSD and a hard drive.

Funding Support: He is supported through this grant in the summer.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

Name	Type of Partner Organization	Location

Name	Type of Partner Organization	Location
Huazhong University of Science and Technology	Academic Institution	Wuhan, China

Full details of organizations that have been involved as partners:

Huazhong University of Science and Technology

Organization Type: Academic Institution
Organization Location: Wuhan, China

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We have collaborated on several research topics and co-authored several papers published.

Have other collaborators or contacts been involved? No

Impacts

What is the impact on the development of the principal discipline(s) of the project?

This project advances the understanding of characteristics of parallel I/O workloads, buffer caching in hybrid storage systems, as well as leveraging emerging memory devices to improve I/O performance for large-scale data-intensive applications.

What is the impact on other disciplines?

Dr. Yifeng Zhu has collaborated with researchers in Marine Science and Earth Sciences to alleviate the I/O bottleneck for their simulations. The climate model developed at UMaine is I/O intensive. We have been collaborating to utilize parallel I/O to speed up the data-intensive science applications. Due to the success of this project, PI have been collaborating with faculties in Earth Science and Bioengineering on big data research projects. Two collaborative proposals have been submitted to NSF.

What is the impact on the development of human resources?

This research project has partially supported two graduate-students, and one post-doctoral research. Our post-doc has received the best paper nominations in HPCA'13.

At the University of Maine, the research findings and concepts have been incorporated into two innovative NSF-funded education programs, directed 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 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 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.

What is the impact on physical resources that form infrastructure?

Due to this research work, we have received equipment donation from LexisNexis, a leading global provider of content-enabled workflow solutions designed specifically for professionals in the legal, risk management, corporate, government, law enforcement, accounting, and academic markets.

What is the impact on institutional resources that form infrastructure?

Due to our research on ultra-large scale data storage systems, our research group along with other researchers in data science and engineering has been selected as one of signature and emerging Areas of excellence in research and education at the University of Maine. This designation is to form strategic and focused planning and resource allocation to preserve UMaine's national stature and impact in Maine.

What is the impact on information resources that form infrastructure?
Nothing to report.

What is the impact on technology transfer?
Nothing to report.

What is the impact on society beyond science and technology?
Nothing to report.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.

Special Requirements

Responses to any special reporting requirements specified in the award terms and conditions, as well as any award specific reporting requirements.
Nothing to report.