2-8-2015

Collaborative Research: HECURA: A New Semantic-Aware Metadata Organization for Improved File-System Performance and Functionality in High-End Computing

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

My Desktop
Prepare & Submit Proposals
Proposal Status
Proposal Functions
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Federal Financial Report History
Financial Functions
Grantee Cash Management Section Contacts
Administration
User Management
Research Administration
Lookup NSF ID

Preview of Award 0937988 - Final Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	0937988
Project Title:	Collaborative Research: HECURA: A New Semantic-Aware Metadata Organization for Improved File-System Performance and Functionality in High-End Computing
PD/PI Name:	Yifeng Zhu, Principal Investigator
Recipient Organization:	University of Maine
Project/Grant Period:	08/15/2009 - 07/31/2014
Reporting Period:	08/01/2013 - 07/31/2014
Submitting Official (if other than PD\PI):	Yifeng Zhu Principal Investigator
Submission Date:	02/08/2015
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 collaborative project is to create a new decentralized semantic-aware metadata organization to speed up the I/O performance. This project exploits semantics of file metadata to improve system scalability, reduce

query latency for complex data queries, and enhance file system functionality.

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

Major Activities: In this project, we have conducted research on the decentralized semantic-aware metadata organization for large-scale storage systems, and make progress as planned. The major activities include designing, implementing, and evaluating the semantic-aware decentralized metadata management schemes to improve file system performance, searchability, and storage efficiency.

Specific Objectives: The specific objectives are: 1) exploit the semantic and scalable nature of the new metadata organization to significantly speed up complex queries and improve file system functionality; 2) Exploit Phase-Change Memory (PCM) to reduce the read I/O latency.

Significant Results: We have exploited semantic correlations among files to create a flat, small, and accurate semantic-aware namespace for each file. The per-file namespace is a flat structure without an internal hierarchy. For a given file, its namespace consists of a certain number of the most closely correlated files. We design an efficient method to identify semantic correlations among files by using a simple and fast LSH-based lookup. For each lookup operation, our design cost-effectively presents users' files that might be of interests. We have implemented it as a middleware that can run on top of most existing file systems, orthogonally to directory trees, to facilitate file lookups. We intend to release it for public use in the near future.

We aim to exploit Phase-Change Memory (PCM) to improve metadata and I/O performance. Better process scalability and less leakage power make Phase-Change Memory (PCM) a promising alternative or supplement to memory and storage systems. The increasing number of cores in the Chip Multi-Processor (CMP) and workloads with large amount date set pose a denser storage challenge. Multiple level cell (MLC) increases storage density and makes PCM more scalable. We have proposed a bit-mapping scheme to improve MLC read latency with minimal hardware overhead. Taking 2-bit MLC as an example, MCWM stores the first half of each cache line at most-significant-bits (MSBs) of MLC cells, and the second half at least-significant-bits (LSBs). Simulation results show that our design can successfully reduce the read latency of standard MLC PCM baseline by 27.5% on average for memory I/O intensive workload.

In addition, we also analysis the I/O overwrite rates. We found that in metadata traffic, there is a significant fraction of block overwrites in both server and desktop workloads. Metadata overwrites are particularly severe. A critical challenge of using PCM to speed up metadata accesses is the reliability (write endurance) issue of PCM.

Key outcomes or Other achievements: This project has led to:

- A number of effective and efficient semantic-aware namespace and metadata management approaches that can be used to support fast and accurate file search for ultra large file systems.
- Exploration of emerging none-volatile to speed up I/O performance.
- A few top tier publications.
- Training of one Postdoc researcher and one female PhD student.

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

This project has trained one Postdoc researcher and one female Ph.D. student.

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

We have published a few papers in top conferences and journals, such as IEEE Transactions on Parallel and Distributed Systems.

Products

Books

Book Chapters

Conference Papers and Presentations

Jianhui Yue (2015). *Reducing Read Latency in MLC PCM*. DATE 2015. . Status = UNDER_REVIEW; Acknowledgement of Federal Support = Yes

Inventions

Journals

Y. Hua, H. Jiang, Y. Zhu, D. Feng, and L. Xu (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

Yu Mao, Jiguang Wan, Yifeng Zhu, and Changsheng Xie (2014). A New Parity-Based Migration Method to Expand RAID-5. *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS*. 25 (8), 1945. Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Licenses

Other Products

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations

Websites

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Zhu, Yifeng	PD/PI	0
Yue, Jianhui	Postdoctoral (scholar, fellow or other postdoctoral position)	5
Rahman, Tania	Graduate Student (research assistant)	6

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: 0

Contribution to the Project: Supervised graduate students and post-doc research fellows; Integrated research results into courses PI taught;

Funding Support: No

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: 5

Contribution to the Project: Work on the research topic of speeding up I/O write performance

Funding Support: Support from this grant

International Collaboration: No
International Travel: No

Tania Rahman
Email: tania.rahman@maine.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 6

Contribution to the Project: She was working on the analysis of the I/O overwrite rate of metadata traffic in typical server workloads.

Funding Support: Partially supported by this grant.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?

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

Full details of organizations that have been involved as partners:

Huazhong University of Science and Technology

Organization Type: Academic Institution
Organization Location: Wuhan, Hubei, China

Partner's Contribution to the Project:
Collaborative Research

More Detail on Partner and Contribution: We collaborate on large-scale metadata management project and co-authored several papers published in premier conference.

What other collaborators or contacts have been involved?

NO

Impacts

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

Storing metadata on emerging non-volatile memory devices, such as PCM, can significantly speed up the I/O speed. We have studies two important issues. (1) How to speed up the read and write operations of PCM? (2) Is the write endurance issue a significant concern?

To improve the write performance of PCM, we have proposed 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 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, the proposed two-stage-write shows 16.5% latency reduction and 9.2% performance improvement over Flip-N-Write.

We also propose a novel scheme to improve the read performance of MLC PCM. Multi-level-cell (MLC) PCM achieves higher storage density than single-level-cell (SLC) at the cost of inferior write and read performance. Slow read latency can become a performance limiter in MLC, particularly under read-intensive workloads. We designed a new and simple bit mapping scheme, called MCWM, which takes advantage of the latency difference between reading most-significant-bits (MSBs) and least-significant-bits (LSBs) of a MLC cell to improve read performance. At the bit level, MCWM strips all bits of a cache line among MLC cells of a PCM line. Taking 2-bit MLC for example, MCWM maps the first half of a cache line to MSBs of MLC cells, and maps the second half to LSBs. This simple mapping strategy fully leverages the distribution regularity of critical words. For all benchmarks studied, on average 87% of critical words are located within the first half of a cache line. On a cache miss, MCWM allows the critical word of the missed cache line to be fetched from multi-level-cell (MLC) at a speed as fast as single-level-cell (SLC), thus reducing the processor stall time.

We also found that the write endurance of PCM is a critical concern for storage workloads. Endurance issue can cause SCM fail very soon. There is a significant fraction of block overwrites in both server and desktop workloads. We studied five enterprise storage workloads. We found that on average 1% of the most written blocks contribute to 34 % of the total overwrites in server traces. In another word, a small fraction of blocks contribute to a large percentage of overwrites. This observation is consistent with findings made by other researchers. If each PCM cell can endure 10^7 overwrites (10^6 reported in and 10^7 reported in and the workloads continuously repeat, the lifetime of SCM can be as short as 33 days in the build-server workload (BS) and 53 days in the exchange-server workload (Exch). Such a short lifespan makes PCM impractical to be used as persistent storage.

What is the impact on other disciplines?

This project has strengthened the research collaborations on campus at the University of Maine. Under the lead of the PI of this project, data science and engineering has been listed as one of the signature and emerging research areas at

the University of Maine. The University will place more emphasis in this research direction and further strengthen interdisciplinary collaborations on data science and engineering.

What is the impact on the development of human resources?

This project has trained one Postdoc researcher and one female Ph.D. student.

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

What is the impact on institutional resources that form infrastructure?

University of Maine has listed data science and engineering as one of the signature and emerging research areas. Data Science and Engineering, leveraging UMaine strengths in data science and engineering, and data-sensitive science areas by applying data-centric methods to issues relevant to Maine's interests and natural and economic sustainability. DSE brings together computer scientists, mathematicians, statisticians and engineers with domain scientists to address critical challenges of capturing, storing, managing, sharing, and analyzing massive data sets for new scientific discoveries and insights.

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.