Department of Energy Project ER25739

Final Report

*“QoS-Enabled, High-performance Storage Systems for
Data-Intensive Scientific Computing”*

Institution	Florida International University
Principal Investigator	Raju Rangaswami
PI Contact Information	Florida International University University Park, ECS 310 11200 SW 8th Street Miami FL 33199 Tel: (305)-348-6230, Fax: (305)-348-3549 raju@cs.fiu.edu

1 Final Project Report

1.1 Summary of Project Accomplishments

This project’s work resulted in the following research projects:

- BORG: Block-reORGanization for Self-optimizing Storage Systems
- ABLE: Active Block Layer Extensions
- EXCES: EXternal Caching in Energy-Saving Storage Systems
- GRIO: Guaranteed-Rate I/O Scheduler

These projects together help in substantially advancing the over-arching project goal of developing “QoS-Enabled, High-Performance Storage Systems”.

1.2 Project Findings

We report our findings from each of the sub-projects below:

BORG

We reported our findings on the BORG project in a paper to appear in USENIX FAST 2009. BORG is a block reorganization layer situated between a file system (or other storage abstraction) and a disk drive. BORG constructs an optimized layout for a subset of the blocks on the device which comprise the current active working-set data used by the workload, using a greedy algorithm applied to a block trace gathered from the system in use. BORG then implements the layout in a separate area of the disk (OPT), and uses an map and heuristics to decide whether I/Os should be serviced from OPT or from their default file system locations. Additionally, writes are also absorbed in the OPT space. BORG improves performance for some workloads, particularly workloads that have a repetitious element (booting a VM or a large app) or workloads that are write intensive, but can degrade performance when a read-mostly workload changes its on-disk working set faster than BORG can react (on the order of hours). BORG incurs significant computational overhead for training (but there is substantial room for improvement to the current implementation) but the memory overhead is reasonable and the data copying can be done as a background activity.

ABLE

An overview and case-making paper for the ABLE project was published at the

IEEE SPEED 2008 workshop in February, 2008. Since then, we have developed a more extensive version of the ABLE block-layer Linux kernel infrastructure and evaluated the effectiveness of such an infrastructure in simplifying the development of several real-world storage self-management extensions. We also developed and extended the theory of block layer extensions that allows the reasoning of multiple block layer extensions when operating simultaneously within a single system. All of this work is currently under review for conference publication.

EXCES

We reported our findings on the EXCES project in a paper that was published at the IEEE Conference on High-Performance Computer Architecture, 2008. We designed and implemented EXCES, an external caching system that employs prefetching, caching, and buffering of disk data for reducing disk activity and conserving energy. EXCES addresses important questions related to external caching, including the estimation of future data popularity, I/O indirection, continuous reconfiguration of the ECD contents, and data consistency. We evaluated EXCES with both micro- and macro- benchmarks that address idle, I/O intensive, and real-world workloads. Overall system energy savings was found to lie in the modest 2-14% range, depending on the workload, in somewhat of a contrast to the higher values predicted by earlier studies.

GRIIO

The I/O scheduler implementation in today's operating systems lacks quality of service. Specifically, I/O scheduling policies implemented inside the kernel do not guarantee the rate at which I/O will be serviced to any process. The efficiency of the I/O request to be completed thus depends on the various other factors like disk seek time, total IO pending to be serviced, active IO scheduler policy. Applications like the continuous media playback require their requests to be processed with a guaranteed rate to ensure correct behavior. Similarly, virtual machine monitors must guarantee that individual VMs are provided with a certain percentage of disk I/O bandwidth. We architected and implemented the Guaranteed Rate IO scheduler (GRIIO) to improve I/O quality of service. GRIIO refers to a guarantee made by the system to a user process that the process will receive data from a system resource at a predefined rate regardless of any other activity on the system.

1.3 Students Supported by This Award

Graduate Students

- Jorge Guerra (Hispanic minority Ph. D. student)
- Ricardo Koller (Hispanic minority Ph. D. student)
- Luis Useche (Hispanic minority Ph. D. student)

1.4 Publications

The following technical articles were produced from the effort in this project:

1. “BORG: Block-reORGanization for Self-Optimizing Storage Systems”
Medha Bhadkamkar, Jorge Guerra, Luis Useche, Sam Burnett, Jason Lip-
tak, Raju Rangaswami, and Vagelis Hristidis.
Proceedings of the File and Storage Technologies (FAST), February 2009.
2. “The Case for Active Block Layer Extensions”
Jorge Guerra, Luis Useche, Medha Bhadkamkar, Ricardo Koller, and Raju
Rangaswami.
ACM Operating Systems Review, 42(6), October, 2008.
3. “EXCES: EXternal Caching in Energy Saving Storage Systems”
Luis Useche, Jorge Guerra, Medha Bhadkamkar, Mauricio Alarcon, and
Raju Rangaswami.
*Proc. of IEEE International Symposium on High-Performance Computer
Architecture (HPCA)*, February, 2008.
4. “Workload-based Generation of Admin Hints for Optimizing Database
Storage Utilization”
Kaushik Dutta, Raju Rangaswami, and Sajib Kundu.
ACM Transactions on Storage, 3(4), February, 2008.
5. “The Case for Active Block Layer Extensions”
Jorge Guerra, Luis Useche, Medha Bhadkamkar, Ricardo Koller, and Raju
Rangaswami.
*Proc. of the IEEE International Workshop on Storage and I/O Virtualiza-
tion, Performance, Energy, Evaluation and Dependability (SPEED)*, held
in conjunction with *IEEE HPCA*, February, 2008.

6. “2LP: A Double Lazy XML Parser”
Fernando Farfan, Vagelis Hristidis, and Raju Rangaswami. *Under peer review, submitted to: Elsevier Journal of Information Systems*, 2008.

1.5 Talks

The following public/industry talks provided a medium for the dissemination of research results related to this project:

1. “BORG: Block-reORGanization for Self-Optimizing Storage Systems”
Medha Bhadkamkar.
Proceedings of the File and Storage Technologies (FAST), February 2009.
2. “New Directions in Self-Managing Storage Systems ”
Raju Rangaswami.
IBM India Research Laboratory, New Delhi, India, May 2008.
3. “Active Block Layer Extensions”
Raju Rangaswami.
VMWare, Inc., February, 2008.
4. “The Case for Active Block Layer Extensions”
Jorge Guerra.
IEEE International Workshop on Storage and I/O Virtualization, Performance, Energy, Evaluation and Dependability (SPEED), February, 2008.
5. “EXCES: EXternal Caching in Energy Saving Storage Systems”
Luis Useche.
IEEE International Symposium on High-Performance Computer Architecture (HPCA), February, 2008.

1.6 Website

Technical articles related to the project can be accessed at :
<http://www.cs.fiu.edu/~raju/WWW/publications/>