論文 / 著書情報
Article / Book Information

Title	A Prototype of Power Saving Storage Method RAPoSDA
Author	Satoshi Hikida, Hiroki Oguri, Haruo Yokota
Journal/Book name	Proc. of SRDS'14, , , pp. 339-340
Issue date	2014, 10
DOI	http://dx.doi.org/10.1109/SRDS.2014.64
URL	http://www.ieee.org/index.html
Copyright	(c)2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

A Prototype of Power Saving Storage Method RAPoSDA

Satoshi HIKIDA
Department of Computer Science
Tokyo Institute of Technology
Japan
Email: hikida@de.cs.titech.ac.jp

Hiroki OGURI
Department of Computer Science
Tokyo Institute of Technology
Japan
Email: oguri.h.aa@m.titech.ac.jp

Haruo YOKOTA
Department of Computer Science
Tokyo Institute of Technology
Japan
Email: yokota@cs.titech.ac.jp

***Abstract*—Because of the development of information technologies and the pervasiveness of cloud services, Reducing power consumption in large scale storage systems becomes a very important issue. Recently, we have proposed a method to solve the problem called Replica Assisted Power Saving Disk Array (RAPoSDA) and verified its effectiveness. RAPoSDA utilizes a primary backup configuration on both cache memories and disk drives to ensure system reliability and it dynamically controls the timing and targeting of disk access based on individual disk rotation states. Until now, we have evaluated the effectiveness of RAPoSDA by using simulation program which we had developed. The evaluation shows that RAPoSDA achieves significantly reducing power consumption. However, we have not evaluated it effectiveness on the practical environment. In this paper, we introduce the prototype system of RAPoSDA to evaluate the power saving capability of RAPoSDA.**

***Keywords*—large-scale storage, power reduction, reliability**

I. Introduction

Nowadays the storage systems owned enterprises and data centers has become large-scale to preserve a number of digital data. Therefore, reducing the power consumption of these storage systems are very important issue. There are several researches tackle this problem, especially MAID[1] is one of the most known the proposal to reduce the power consumption of large scale storage systems. MAID separates its disk drives to two types of disk group. One called cache disks and the another called data disks. The data accessed frequently are stored to few cache disks and then decreasing the access frequency to the many other data disks. Therefore these data disks can sleep long enough time to reduce the total power consumption of storage system. However, MAID does not consider to maintain the reliability and the fixed number cache disks may become the performance bottleneck. Whereas, we have proposed a method to solve this problem called Replica Assisted Power Saving Disk Array (RAPoSDA)[2] and verified its effectiveness. RAPoSDA utilizes a primary backup configuration on both cache memories and disk drives to ensure system reliability and it dynamically controls the timing and targeting of disk access based on individual disk rotation states. Until now, we have evaluated the effectiveness of RAPoSDA by using simulation program which we had developed. The evaluation shows that RAPoSDA achieves significantly reducing power consumption. However, we have not evaluated it effectiveness on the practical environment. In this paper, we introduce the prototype system of RAPoSDA and describe its configuration. We plan to evaluate the effectiveness of RAPoSDA in practical environment and compare with MAID using this prototype system.

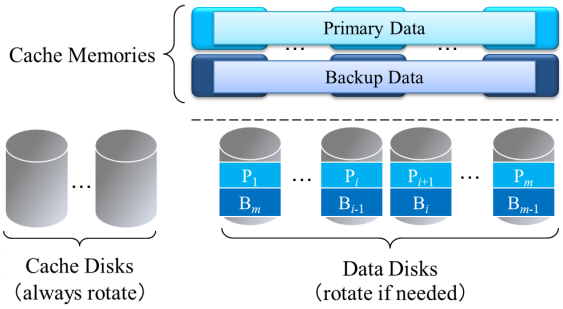

Fig. 1. Configuration of RAPoSDA

II. Overview of RAPoSDA

Fig. 1 shows the configuration of RAPoSDA. From this figure, RAPoSDA consists of several cache memories and disk drives. In particular, disk drives are separated into two types of disks, one called cache disks and the other called data disks. And then RAPoSDA use a write cache memory to temporarily maintain data. It is important to maintain reliability when the data are in the volatile cache memory, so we provide the cache memory with a primary backup configuration that corresponds to the data disks and assume each cache memories are connected to corresponding UPSs. In addition, RAPoSDA can use some disks as cache disks like a MAID[1] to provide larger read cache spaces.

We adopt the chained declustering [3] method for data placement on data disks as a primary backup configuration that tolerates disk failures. Chained declustering is a simple but effective strategy for data placement that provides good reliability and accessibility when the backup data is logically located in the disk next to the primary. Cache memories are dedicated to corresponding to the primary and backup of data. We assume each cache memory connects to the corresponding power supply which assured to avoid power accident by using uninterruptible power supply(UPS) individually.

In RAPoSDA, the data were replicated to two copies (primary and backup) on both cache memories and data disks. At first, a primary data is assigned to a corresponding data

disk and then written to the primary area of a cache memory which is binding to a disk group(DG) which encloses the data disk. The backup data of the primary data is written to the backup area of the another cache memory which assigned randomly. Meanwhile, because of data disks employees chained declustering[3] data placement policy, a backup data which corresponding to the i-th disk's primary data is stored on $((i+1) \bmod N_{DD})$-th disk, where N_{DD} represents a number of data disks.

A. *Read/write behaviour of RAPoSDA*

To handle write request, the data are initially written into both the primary and backup layer of the cache memory. The written data are gathered in a corresponding buffer location in the cache memory of each individual disk that is responsible for storing the data. Buffered data are written onto their corresponding data disks when the amount of buffered data on the cache memory exceeds a predefined threshold.

read requests initially check the existence of data in the cache memory, followed by cache disks. The cache memory has primary and backup layers, and both layers are searched for data. If the target data are in the cache memory or cache disks, the data are returned without accessing the data disks. If the data do not exist in the cache memory or cache disk, the data are read from a data disk.

III. Effectiveness of RAPoSDA

Recently, we demonstrated the effectiveness of RAPoSDA in our work [2]. In that work, we compared the performance and power saving effects of RAPoSDA with modified MAID in simulations with different ratios of read/write requests in the workloads. The original version of MAID had no cache memory and no replication mechanism, so we added them to MAID to make a fair comparison. The experimental results showed that RAPoSDA and the modified MAID provide reduced power consumption compared with a simple disk array with no power saving mechanism. However, RAPoSDA was superior to the modified MAID. From the performance perspective, the simulation results showed that the average response time of RAPoSDA was shorter than that of the modified MAID. Thus, consideration of individual disk rotation is an effective method for reducing the power consumption of storage systems, while maintaining good performance.

IV. Prototype system

In this section, we describe a prototype storage system for evaluation of our proposal called RAPoSDA. Figure 2 depicts the configuration of our prototype system. The prototype system consists of a PC-based controller, host bus adapter(HBA), and power measuring instrument.

The controller is built on Linux kernel 2.6.32-5-amd64 and implemented as a key value store written in Java. In addition, to avoid the file system's cache mechanism we wrote the C++ native codes to direct disk accesses using open(1) system call with O_DIRECT flag and our prototype call this module via JNI mechanism. As host bus adapter, we employee *High Point Data Center 7280 which can attach up to 32 SATA HDD to one PC via PCI Express interface. The power measuring instrument is* HIOKI Memory Hilogger which has

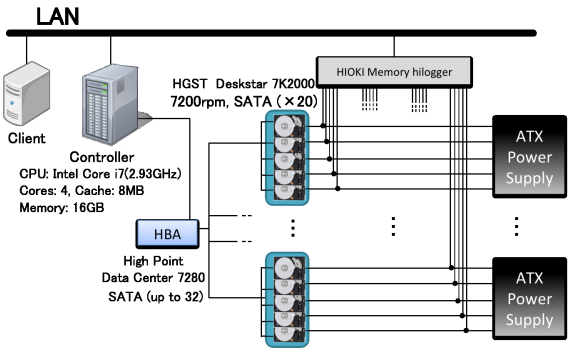

Fig. 2. The configuration of prototype system

90 channel to measure the currents and volts. we use 4 channel to measure one HDD power consumption. Two of them probe 12V line and grand line, the other probe 5V line and grand line, respectively.

Incidentally, we assume the buffers (or DRAM memory modules) of RAPoSDA are connected to the power supply independently each other. However, it is hard to organize such a configuration on the actual commodity PC hardware. We think there are few affect whether each buffer connects to corresponding power supply individually, the prototype has a power supply. Although the buffer share the power supply, we treat the buffer as several number of buffers in logically.

V. Conclusion

In this paper, we describe the prototype system for evaluation of our proposal that called RAPoSDA in practical environment. the protype can configuration other type of storage configuration, then we plan to compare with MAID. we have evaluated and verified the effectiveness of RAPoSDA by using simulation program base experiment. However, we do not its actual effectiveness in practical environment.

In the future we plan more experimental evaluation using real workloads to verify the effectiveness of RAPoSDA capability precisely. In addition we plan to compare the prototype based experiment result with the result of simulation program based evaluation.

Acknowledgment

This work is partly supported by Grants-in-Aid for Scientific Research from Japan Science and Technology Agency (A) (#22240005, #25240014).

References

[1] D. Colarelli and D. Grunwald, "Massive arrays of idle disks for storage archives," in *Supercomputing '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*. Los Alamitos, CA, USA: IEEE Computer Society Press, 2002, pp. 1–11.

[2] S. Hikida, H. H. Le, and H. Yokota, "A power saving storage method that considers individual disk rotation," in *The 17th International Conference on Database Systems for Advanced Applications (DASFAA)*, vol. 7239/2010, April 2012, pp. 138–149.

[3] H.-I. Hsiao and D. J. DeWitt, "Chained declustering: A new availability strategy for multiprocessor database machines," in *Proceedings of the Sixth International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1990, pp. 456–465.