論文 / 著書情報
Article / Book Information

題目(和文)	高可用かつ高信頼なストレージシステムの省電力化に向けたデータ管理に関する研究
Title(English)	A Study on Data Management for Saving Power of Highly Available and Reliable Storage Systems
著者(和文)	引田諭之
Author(English)	Hikida Satoshi
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第11231号, 授与年月日:2019年6月30日, 学位の種別:課程博士, 審査員:横田 治夫,佐伯 元司,権藤 克彦,吉瀬 謙二,金子 晴彦
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第11231号, Conferred date:2019/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

A Study on Data Management for Saving Power of Highly Available and Reliable Storage Systems

by
Satoshi Hikida

Submitted to the Department of Computer Science, Graduate School of Information Science and Technology in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science
at the
TOKYO INSTITUTE OF TECHNOLOGY
May 2019.

The use of data in modern society is indispensable in various fields. In recent years, the amount of data generated and stored is explosively increasing. For instance, the multimedia data such as videos and images generated by humans or surveillance cameras, and sensor data generated by IoT devices has become stored on large-scale storage systems in data centers to get business insights.

In order to manage such a huge amount of data with high availability and high reliability, it is common in the storage systems to place the replicated data on multiple HDDs to make the data redundancy. Therefore, such a storage system requires a large amount of HDDs, and there is a concern about an increase in the power consumption of them. In fact, among the IT equipment in the data center, the power consumption of the storage system is the second largest after the server, and it is increasing year by year. Therefore, reducing the power consumption of such a large-scale storage system is an important issue from the economic and environmental perspective.

In the case of the HDD-based storage system, the power consumption of the entire storage system can be reduced by stopping the rotation of the disk and putting it into

the standby state. However, if spin-up or spin-down is performed frequently, there is a problem that the power consumption increases rather than in case of always spinning the disk and the response performance deteriorates significantly. To achieve the power saving effect, the HDD is needed to stay the standby state longer than the Break-Even time. Break-Even time is a time of how much power savings can be obtained by stopping the HDD.

This dissertation addresses the problem of the power consumption of highly available and highly reliable HDD-based storage systems. This thesis proposes two novel energy-efficient disk access control techniques for the HDD-based storage system, which has primary-backup configuration to ensure availability and reliability.

First, this dissertation proposes *Replica Assisted Power Saving Disk Array* (RAPoSDA), which is a storage system power saving technique that takes into account the rotation of individual HDDs to control the Read/Write in a disk array with Primary/Backup configuration. In RAPoSDA, to ensure the system reliability, the data on the HDD is configured in a Primary/Backup configuration in a Chained Declustering manner. In addition, to ensure the Read/Write performance and long enough standby period, a non-volatile buffer is used. And the data in the buffer are also configured in a Primary/Backup configuration. Based on this, RAPoSDA has *Group Write* mechanism that Writes-Back the primary and backup data simultaneously to the identical HDD when the buffer threshold is exceeded, and *Selective Read* mechanism that preferentially reads from the rotating HDD. By introducing these mechanisms, the standby period can be assured more than before.

This dissertation evaluates the effectiveness of the RAPoSDA by several approaches. At first, I made an approximate formula using parameters such as the disk rotation probability and the buffer writability to compare the power consumption of RAPoSDA with Normal which does not stop the HDD rotation, and MAID (Massive Arrays of Idle Disks) which is a conventional power-saving technique. The result showed that the power saving effect of the RAPoSDA is 35% higher than that of MAID. Furthermore, since the impact of the workload could not be evaluated using the approximate formula, I developed a simulator to evaluate the configuration of RAPoSDA, Normal, and MAID. And the evaluation was performed with varying the number of HDDs from 32 to 1024 using the workload with varying Read/Write ratio. As a result of the evaluation, it was confirmed that while MAID has a power saving effect of less than

80% for Normal, RAPOSDA has 90% or more, thus, RAPOSDA has a higher power saving effect than that of MAID even if the number of HDDs and the Read/Write ratio are changed. The evaluation also evaluated the response time to the workload. The result showed that the response time of RAPoSDA was smaller than MAID even if the read/write ratio and the number of HDDs were changed, and that the average latency of RAPoSDA could be reduced by 76% or more than MAID.

In addition, in this dissertation, I constructed a prototype system of RAPoSDA using an actual HDD and verified its applicability to real environments. A prototype experiment using actual 20 HDDs and power measurement equipment confirmed that RAPoSDA achieved higher power savings and response time than that of MAID. In terms of power saving, RAPoSDA has been demonstrated that power consumption can be reduced by 36% compared with Normal.

In the evaluation of the system reliability, this dissertation calculates the mean time to data loss (MTTDL) of RAPoSDA and MAID from the mean time to failure (MTTF) of the system components. The evaluation result showed that the difference of the MTTDL is very small and it can be negligible. The system configuration cost was also evaluated, showing that the initial cost of RAPoSDA, which depends on the system configuration, can be amortized in about two months to two and a half years by reducing the power cost due to the power saving.

Next, to address the problem of increasing disk access frequency as increases the write operation of highly redundant storage system, this dissertation proposes two data placement policy called *Disk Group Aggregation* (DGA) and *Cache Striping* (CS), which groups a buffer and some corresponding data disks, and placed the replica data across the groups to reduce the power consumption of such a highly redundant storage system.

In DGA, primary data is placed in the same group of buffers, and in CS, primary data is placed in a buffer of a different group. In both methods, replica data on the buffer is logically adjacent to the buffer, and replica data on the disk is logically placed on the adjacent disk. They change the timing exceeding the capacity threshold of the buffer for each replica, and when writing the data to the disk, selecting the target disk in consideration of the rotation status suppresses the frequency of disk accesses and reduces the frequency of redundancy of the storage system Suppress power consumption. Through simulation evaluation, DGA is suitable for workload with biased

data access, CS confirmed that data access is suitable for random workload.

In addition, I propose two buffer flushing methods, WithAllSpins and SpinupEE, as power saving buffer management. WithAllSpins flushes the buffer data not only on the disk to be written when the buffer capacity threshold is exceeded, but also on all the disks being rotated at that point, thereby avoiding spin-up of the disk, secure a lot of free space. In addition, SpinupEE estimates the time from the past history to the next buffer overflow, calculates the power efficiency by using the time up to the next buffer overflow estimated as the data on the buffer, and calculates the power efficiency spin-up and flush the buffer for that disk as well.

By simulation evaluation using synthetic workload, I confirmed that both WithAll-Spins and SpinupEE reduced the buffer overflow count and the spin-up/spin-down count more than RAPoSDA's buffer flushing approach.

These methods proposed in this research can be applied singularly or in combination to the large scale storage system being used today, contributing to power saving of the storage system.