Energy aware RAID Configuration for Large Storage Systems

Norifumi Nishikawa
Institute of Industrial Science
The University of Tokyo
norifumi@tkl.iis.u-tokyo.ac.jp

Miyuki Nakano
Institute of Industrial Science
The University of Tokyo
miyuki@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
Institute of Industrial Science
The University of Tokyo
kitsure@tkl.iis.u-tokyo.ac.jp

Abstract

Power consumption of storage at data centers is increasing rapidly. Large storage facilities have various RAID configurations incorporating different RAID levels, numbers of drives, and media types. Nevertheless, few discussions of RAID configurations have been pursued from an energy saving perspective. We first investigate how different RAID configurations affect not only application performance but also power consumption of storage installations. We then present simulation results of power consumption and application performance conducted with various RAID configurations. Results show that a RAID configuration strongly impacts energy conservation based on application I/O features.

1. Introduction

Data stored in data centers continues to accumulate rapidly. This rapid data growth requires more storage capacity: the annual growth rate of external storage shipments is 57% [1]. The compounded annual growth rate of storage power consumption is 19%, which is much higher than expenditures for other IT equipment [1]. Storage energy conservation, specifically limitation of consumption, will be a crucial goal at data centers.

A large storage system comprises hundreds of drives. It can be configured as multiple RAID groups with different RAID levels or different numbers of drives [10]. A RAID group tends to be configured as RAID 5 or 6, with dozens of drives arranged for performance, capacity, and efficiency. Such RAID groups, however, might not be efficient for energy saving because the unit of power control becomes large. Aiming to realize storage energy saving, it is important to use various RAID levels, or to select an appropriate number of drives of the RAID group according to application I/O behaviors.

Many energy-saving storage methods have been proposed [2–8]. However, few studies have examined the influence that the number of drives or a RAID level has upon storage power consumption.

As described in this paper, we investigate the effects of different RAID configurations from the perspective of energy conservation. Here, RAID configurations are designated by the number of drives, the RAID level, and media types.

First, we consider the number of drives in a RAID group. The energy-saving overhead of a RAID group consisting of dozens of drives is quite large (e.g., the waiting time to spin up the RAID group is longer than 1 min). An appropriate selection of the drive number in a RAID group might be effective for energy-saving storage uses.

Second, we consider a RAID level. Popular RAID levels are RAID 5 and 6. RAID 5 (including RAID 6) is suitable for performance or capacity efficiency. However, RAID 5 might not be a good energy-saving configuration because all drives in a RAID group are accessed during reading or writing data. Nevertheless, drives for redundant data of a RAID 4 or 0+1 are not accessed when reading data. We were able to turn off these drives to conserve power used for storage. Therefore, we study an individual drive's power control in a RAID group. Usually, storage in a RAID group turns all drives on and off. A hard disk drive (HDD) has a power saving function. The waiting time to spin up HDD is only a few seconds, which is much shorter than that of the RAID group. Therefore, we can apply the HDD power saving function to short I/O intervals so that the RAID group power saving functions are not applicable.

Third, we also consider drives of the media type. Today, solid state disks (SSDs) are likely to be an alternative low-energy device to limit storage power consumption. However, fewer studies of SSD have been done from a perspective of runtime energy-saving. The energy and performance characteristics of SSD differ greatly from those of HDD because SSD has no mechanical parts such as HDD heads and arms. The suitable RAID configuration for SSD differs from traditional RAID configurations.

Through our measurements, we use database benchmark programs (TPC–C and TPC–H), which are typical I/O intensive applications at data centers. We simulate the power consumption and performance of various RAID configurations with real I/O trace of TPC–C and TPC–H. Simulations show that, from application I/O features, a RAID configuration has a strong impact on energy saving.

978-1-4577-1221-0/11/$26.00 ©2011 IEEE

Section II provides explanations of work related to disk storage energy saving. Section III describes the power consumption characteristics of large storage. Section IV presents the energy saving potential of the RAID configuration. In section V, we describe the evaluation results of our approach. Finally, we conclude this paper.

2. Related Works

Many energy-saving storage methods have been reported in the relevant literature. Earlier reports [2–4] presented approaches that move a status of parity or mirror HDD to low-power mode. Another [5] explained an approach that controls the I/O schedules to enlarge I/O intervals using a storage cache. Other research efforts [6, 7] have examined approaches to control the HDD rotation speed according to I/O traffic. Another paper [8] explained an approach to write data onto an appropriate HDD according to its I/O pattern. These energy-saving approaches [2,3,5–8], however, do not consider the effect of changing the number of drives, RAID level, or media type in a RAID group for storage energy savings. A system that changes the number of disks in a RAID group dynamically according to load of applications has been proposed from one study [4], which did not evaluate the power and performance of a different RAID configuration without changing the total number of disks.

3. Power Consumption Characteristics

3.1. Hard Disk

We measured HDD (750 GB, SATA 7200 rpm, Barracuda ES ST3750640NS; Seagate Technology LLC) power consumption using a digital power meter. The I/O pattern is a combination of random read and write, with I/O size of 8 KB. The HDD power consumption was increased according to the IOPS increment. The maximum power consumption of the HDD is about 14 VA, which is 40% higher than the power consumption of an idle state.

The HDD has a power saving function called the standby state, which stops the plate rotation and parks all heads. The power consumption of the standby state is approximately 1.5 VA. Spinning up the HDD requires more than 8 s and 186 J.

3.2. Solid State Disk

Table 1 presents power consumption data of an SATA solid state drive (X25-E Extreme; Intel Corp.) [9]. As Table 1 shows, the SSD power consumption at an active state is 2.4 W; that for an idle state is 0.06 W. The active state SSD power consumption is about one-sixth of those of HDD. The power consumption of idle state SSD is approximately 1/150 of those of HDD.

Table 1. Power Consumption of SSD

Status	Power consumption
Active*1	2.4 W
Idle (DIPM)	0.06 W

*1: Active energy is measured on an IOMeter workload of full bandwidth 64 K sequential writes with queue depth 1.

3.3. RAID Group

We measured power consumption characteristics of a RAID group in an actual storage system (Adaptive Modular Storage 2500; Hitachi, Ltd.) using a digital power meter. The RAID group has 15 HDDs. The RAID level is 6 (13D+2P). The HDD model is the same as described at section 3.1. The I/O pattern is the same pattern described at section 3.1. Power consumption of the RAID group increases slightly from idle status in accordance with the increase of IOPS. The maximum power consumption is about 315 VA (+10.6% from the idle). The RAID group also has a power saving function of two types called a spin down and a power off. The spin down function turns off all HDDs in a RAID group. The power-off function turns off HDDs and a power source in the RAID group. The power consumption is decreased by 57.4% at a spin down status, and by 100.0% at the power-off status. Spinning up the RAID group requires more than a hundred seconds and tens of thousands of Joules.

3.4. Break Even Time

The “break-even time” is the length of idle time for which the energy required for spinning up the HDDs or RAID groups is equal to the energy saved by maintaining the idle state. For reducing the power consumption, the length of the access interval of HDDs or RAID groups must be longer than the break-even time. Table 2 presents the energy-saving method, spin up wait time, and break-even time of an HDD, an SSD, an HDD RAID group, and an SSD RAID group. Here, we assumed that the power consumption and a spin up wait time of the power source of SSD RAID group are both 1/6 of HDD RAID group. Here, 1/6 is a ratio of power consumption of active status of SSD to the active status of HDD.

Table 2. Spin Up Wait Time and Break-even Time

Type	Power Saving Method	Spin up wait time	Break-even Time
HDD	Standby	8 s	25.5 s
SSD	Idle	1 s	1.0 s
RAID group (HDD)	Spin down	16 s	27.0 s
	Power off	69 s	51.0 s
RAID group (SSD)	Spin down	2 s	3.7 s
	Power off	9 s	5.3 s

4. Energy Saving Potential of RAID Group

4.1. Number of Drives in a RAID Group

Today, a RAID group has dozens of drives. Such a large RAID group might have a large overhead for energy saving of storage because dozens of drives are turned on and off simultaneously. A small number of drives in a RAID group such as 5 or 6 drives might reduce the power consumption of the RAID group because the I/O workload of small random read / write does not require much higher throughput than that of the bulk sequential read.

4.2. RAID Level

The access pattern of drives in a RAID group differs among RAID levels. For RAID 4 or 0+1, some drives have only redundant data. These drives provide more energy saving opportunities because they are not accessed during data reads. The RAID levels offer the following energy-saving potential:

RAID 5: RAID 5 is a default RAID level. In RAID 5, data and parity are stored uniformly on all drives; they are accessed during reading and writing data. Therefore, the energy saving potential is low.

RAID 4: In RAID 4, redundant data are stored in only one parity drive. The parity drive is not accessed during reading data. Therefore, we turned off the parity drive for read-only workloads. For writing data, all drives are accessed, so the power saving potential during writing is the same as that of RAID 5.

RAID 0+1: In RAID 0+1, mirror disks are used for storing redundant data. These mirror disks are not accessed if the workload is low and read-only, so the mirror disks could also be turned off for such a workload. Furthermore, the I/O for updating redundant data is less than that for either RAID 5 or RAID 4. Therefore, RAID 0+1 has a greater chance for application as an energy saving function.

4.3. Combined Use of Power Saving Functions of Drives and RAID Groups

We also evaluated the combined use of drive level and RAID group power saving functions. The break-even time of drives is shorter than that for a RAID group. Therefore, the combined use of these power-saving functions might save more energy than applying the RAID level power saving function alone.

4.4. Energy-Saving Potential of SSD

As Table 1 shows, the power consumption of SSD is 1/6 of HDD. Furthermore, as shown in Table 2, a break-even time and spin up wait times of SSD or SSD RAID group are much shorter than those of an HDD or HDD RAID group, which means that SSD has great potential for RAID group energy savings. Furthermore, the short break-even time of SSD might require another energy-saving function that differs from an energy saving function for HDD RAID groups.

5. Evaluation

5.1 Evaluation settings

RAID Configurations: We compared the power consumption and performance of 15 drives with those of 8 RAID groups (15x8) and 5 drives x 24 RAID groups (5x24). We compared the power consumption and performance of RAID level 5, 4, and 0+1 configurations. We used HDD (SATA) and SSD for evaluation.

Power Saving Function: First, we evaluated the RAID group level power saving function. Then we evaluated their combined use with the drive level and RAID group level power-saving functions.

Spin up wait time and Break-even time: We used a spin up wait time and break-even time values in Table 2 for 15x8 configuration. For 5x24, we used 1/3 of the spin up time of 15x8 because the number of drives is 1/3. The break-even time is also 1/3 because the spin up energy is 1/9 (# of disks is 1/3 and spin up time is 1/3) and the saved energy is 1/3 for 5x24.

Application setting: We compared the power consumption and performance of storage systems using TPC–C and TPC–H applications, which are typical database benchmarks. TPC–C is an online transaction benchmark. Its typical I/O pattern is a small random read/write. TPC–H is a decision-support system benchmark; its typical I/O pattern is large bulk of sequential read. Table 3 shows the respective application settings.

Table 3. Application settings

Application	DB Size	DB Buf Size	Conditions
TPC–C	500 GB (# of warehouse is 5000)	25 GB	# of Threads: 1000 Think Time: 0
TPC–H	100GB (SF=1000)	5 GB	Run Query 1 to 22 one by one.

Data Placement: We classified DB files into two groups: cold files and hot files. A cold file can save the RAID group energy consumption when we put the file into the RAID group alone and apply power-saving functions. Other files are hot files. Then we calculated the number of hot RAID groups which store hot files. We calculated the number of the hot RAID group by dividing the sum of IOPS of hot files by a maximum IOPS of a RAID group. Table 4 presents the number of hot RAID groups with each RAID configuration.

Here, the total number of disks for Hot RAID 0+1 groups of 5x24 TPC–C (20x4=80) is smaller

than that of 15x8 (15x7=105). The reason is that the number of I/O to disks for generating redundant data for RAID 0+1 is smaller than that of RAID 4 or 5.

Table 4. Number of Hot RAID groups

Application	# of drives in a RAID group	RAID Level	# of Hot RAID groups
TPC–C	15 (8 RGs)	5, 4, 0+1	7
	5 (24 RGs)	5, 4	21
		0+1	20
TPC–H	15 (8 RGs)	5, 4, 0+1	1
	5 (24 RGs)	5, 4, 0+1	1

5.2. Simulation Methods

Actual I/O trace: We obtained the actual I/O trace by running TPC–C and TPC–H on our storage system (see subsection 3.3), and measured the power consumption and performance. The RAID configuration is RAID 5 with 15 drives. We then simulated the power consumption and performance of other RAID configuration (RAID 4, RAID 0+1, and 5 drives RAID group) using the trace.

Calculation of Power consumption: We calculated the power consumption of RAID groups using equations presented in Table 5. Here, *i* is the number of I/Os made to a storage unit per second.

Table 5. RAID Group Power Consumption

# of Disks	Media Type	Equations	#
15	HDD	$-1.594\times10^{-3}i^2 + 0.036i + 287.5$ ($i<=2000$) $-1.840\times10^{-6}i^2 + 0.094i + 285.4$ ($i>2000$)	1
	SSD	$2.4*15 + 48$ ($i>0$), $0.06*15 + 48$ ($i=0$) Here, 48 is a power consumption of base part. We assumed that the power consumption of the base part is 1/6 of the base part of HDD RAID Group because the power consumption of SSD is 1/6 of HDD.	2
5	HDD	1/3 of Equation #1	3
	SSD	1/3 of Equation #2	4

The power consumption of the power saving functions for 15x8 is shown in 3.3. For 5x24, we also assumed that the power consumptions are 1/3 of 15x8. We used the power consumptions of a HDD and a SSD described at subsections 3.1 and 3.2.

Calculation of Performance: We assumes that a transaction that issues read I/O to drives or RAID groups in energy saving mode is made to wait until the drives or RAID group is spin up. Other transactions are not delayed. Furthermore, we assume that the query is delayed until the drives or RAID group is spin up if a query issues an I/O to drives or RAID groups in energy saving mode. The spin up wait time is presented in Table 2. We assume that the I/O rate does not change even if the response time of drives or the RAID group is short.

5.3. Evaluation Results

Variation of RAID Configurations at HDD: Figures 1(a) and 1(b) show the power consumption and transaction throughput of TPC–C; Figs. 1(c) and 1(d) shows a power consumption and query response time of Query 7 of TPC–H. Here, the power consumption and performance of 15x8 RAID 5 is actual measured values. In this evaluation, we use only a RAID group energy-saving function.

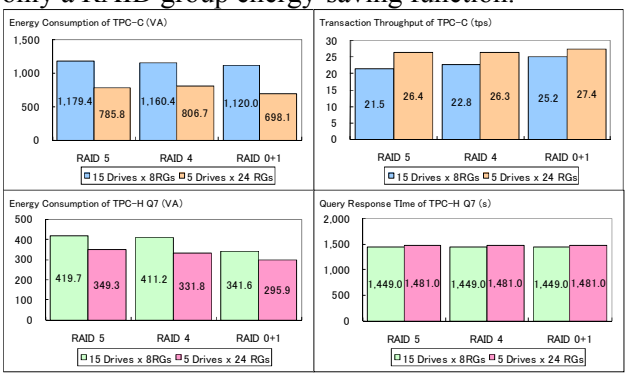

Figure 1. HDD RAID Group Power and Performance.

As shown in Figure 1, for TPC–C, the power consumption of a 5x24 is smaller than that of a 15x8. The 5x24 configuration has three cold RAID groups and only one RAID group was turned on (5 disks) when I/Os were issued to a cold RAID group. However, for 15x8, only one cold RAID group (15 disks) was turned on when I/Os were issued to the cold RAID group. The performance of the 5x24 configuration is better than that of the 15x8 configuration because the break-even time of 5x24 is shorter than 15x8. Therefore we were able to apply the power saving functions more frequently.

For TPC–H, the power consumption of RAID 4, 0+1 is smaller than RAID 5 because I/O pattern of TPC–H is read only and redundant data are not accessed. Therefore we can turn off the drives storing redundant data.

These results demonstrate that a small RAID configuration is effective for energy savings of storage for TPC–C. We can also observe the power consumption reduction of the RAID group for TPC–H in RAID 4 and 0+1 compared with RAID 5. The performance was not decreased.

Variation of RAID Configurations at SSD: Figures 2(a) and 2(b) show the power consumption and transaction throughput of TPC–C, Figures 2(c) and 2(d) show the power consumption and query response time of Query 7 of TPC–H. The RAID configuration and energy saving function are the same as HDD RAID. As Figure 2 shows, we can reduce the power consumption of RAID groups of 5 drives without performance degradation, just as we can for the HDD RAID group. In RAID 4 and 0+1, we can also reduce the power consumption without performance degradation. The reason is the same as the reason for HDD RAID groups. These results show that the RAID level reconfiguration and

number of disks in a RAID group are also effective for reducing the SSD storage power consumption.

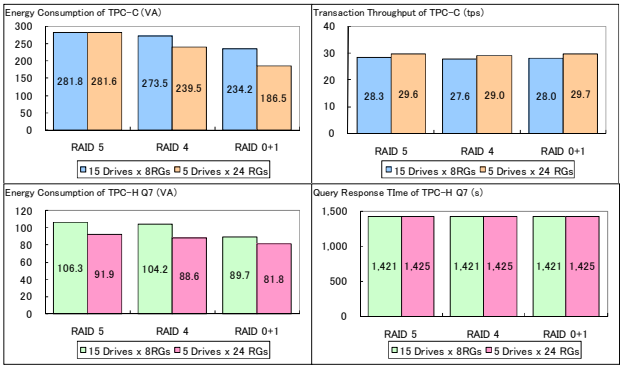

Figure 2. Power and Performance of SSD RAID Group.

Combined Use of Power Saving Functions: We evaluate an effect of an individual drive's power control for cold SSD RAID groups. Figures 3(a) and 3(b) show an energy-saving rate of TPC–C and TPC–H of cold SSD RAID groups. The energy saving function of an individual drive can reduce 10% of the power consumption compared to the energy control of a RAID group. A break-even time of a SSD drive is shorter than that of SSD RAID group. We have much more chance of power saving. Additionally, we cannot reduce the power consumption of cold HDD RAID groups because the break-even time of HDD is much longer than that of SSD. We showed that the combined use of drive's energy saving function and RAID group's energy saving function is one solution to SSD RAID.

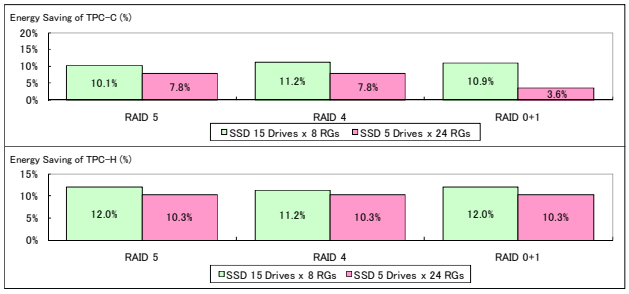

Figure 3. Power Saving Rate of Combined Use of Power Saving Functions.

6. Conclusion and Future Works

We examined the power-saving potential of different RAID configurations to promote energy saving of storage installations, and investigated what combinations of RAID configurations and applications (TPC–C and TPC–H) are effective for energy saving. Simulation results show that a small RAID configuration with RAID 0+1 has more energy-saving potential than a conventional configuration with RAID 5 does. Our quantitative evaluation confirmed that the reconfiguration of the RAID level and number of disks in a RAID group is effective to reduce storage power consumption. We also showed that the energy saving function of considering individual drives presents one solution for SSD RAID group energy savings. Finally, application of I/O features aspects such as small random read/write in TPC–C and large bulk of sequential read in TPC–H show that a RAID configuration can affect energy conservation strongly.

We plan to develop a framework to reduce power consumption by selecting appropriate RAID configurations according to the application I/O behaviors including SSD and HDD hybrid storage. We can then implement RAID configuration tools for energy-aware RAID configuration tools for high performance storages. We can then evaluate those tools and configurations using several workloads.

7. References

[1] B. Patrick et al., "Green Storage II: Metrics and Measurement" http://net.educause.edu/ir/library/pdf/churiedel.pdf. (2010).

[2] E. Pinheiro et al., "Exploiting Redundancy to Conserve Energy in Storage System," Proc. of the Joint International Conference on measurement and Modeling of Computer Systems, 2006.

[3] J. Wang et al. "eRAID: Conserving Energy in Conventional Disk-Based RAID System," IEEE Trans. on Computers, Vol. 57, No. 3, pp. 359-374, 2008.

[4] C. Weddle et al., "PARAID: A Gear-Shifting Energy-Aware RAID," Fifth USENIX Conference on File and Storage Technologies (FAST '07), 2007.

[5] D. Li et al., "EERAID: Energy Efficient Redundant and Inexpensive Disk Array," 11th Workshop on ACM SIGOPS European Workshop, 2004.

[6] S. Grurmurthi et al., "Reducing Disk Energy Consumption in Servers with DRPM," IEEE Computer, Vol. 36, No. 12, pp. 59-66, 2003.

[7] Q. Zhu et al., "Hibernator: Helping Disk Arrays Sleep through the Winter," Proc. Twentieth ACM Symposium on Operating Systems Principles, 2005.

[8] N. Joukov et al., "GreenFS: Making Enterprise Computers Greener by Protecting Them Better," The European Professional Society on Computer Systems (EuroSys), 2008.

[9] "Intel® X25-E Extreme SATA Solid State Drive SSDSA2SH032G1, SSDSA2SH064G1 Product Manual," ftp://download.intel.com/design/flash/NAND/extreme/extreme-sata-ssd-datasheet.pdf, 2009.

[10] D.A. Patterson et al., "A case for redundant arrays of inexpensive disks," Proc. of the 1988 ACM SIGMOD International Conference on Management of Data, 1988

[11] M. Balakrishanan et al., "Differential RAID: rethinking RAID for SSD reliability," Proc. Fifth European Conference on Computer Systems, 2010.